OTTAWA: Optimal TransporT Adaptive Word Aligner for Hallucination and Omission Translation Errors Detection

Chenyang Huang[1†*] **Abbas Ghaddar**[2*] **Ivan Kobyzev**[2]
Mehdi Rezagholizadeh[2] **Osmar R. Zaïane**[1] **Boxing Chen**[2]

[1]Dept. of Computing Science, Alberta Machine Intelligence Institute (Amii), University of Alberta
[2]Huawei Noah's Ark Lab, Montreal Research Center, Canada

{chenyangh,zaiane}@ualberta.ca
{abbas.ghaddar,ivan.kobyzev}@huawei.com
{mehdi.rezagholizadeh,boxing.chen}@huawei.com

Abstract

Recently, there has been considerable attention on detecting hallucinations and omissions in Machine Translation (MT) systems. The two dominant approaches to tackle this task involve analyzing the MT system's *internal* states or relying on the output of *external* tools, such as sentence similarity or MT quality estimators. In this work, we introduce OTTAWA, a novel Optimal Transport (OT)-based word aligner specifically designed to enhance the detection of hallucinations and omissions in MT systems. Our approach explicitly models the missing alignments by introducing a "null" vector, for which we propose a novel one-side constrained OT setting to allow an adaptive null alignment. Our approach yields competitive results compared to state-of-the-art methods across 18 language pairs on the HalOmi benchmark. In addition, it shows promising features, such as the ability to distinguish between both error types and perform word-level detection without accessing the MT system's internal states.[1]

1 Introduction

Concerns regarding hallucination (as known as fabrication) of Machine Translation (MT) systems (Raunak et al., 2021; Müller and Sennrich, 2021; Guerreiro et al., 2023a) have led to considerable efforts towards the creation of diagnostic datasets (Zhou et al., 2021; Guerreiro et al., 2023c; Dale et al., 2023b) and developing detection methods (Guerreiro et al., 2023b; Dale et al., 2023a). In addition to hallucination errors, which occur when words in the translation are detached from the source sequence, addressing omission errors is also crucial—these are cases where words from the source sequence are absent in the translation.

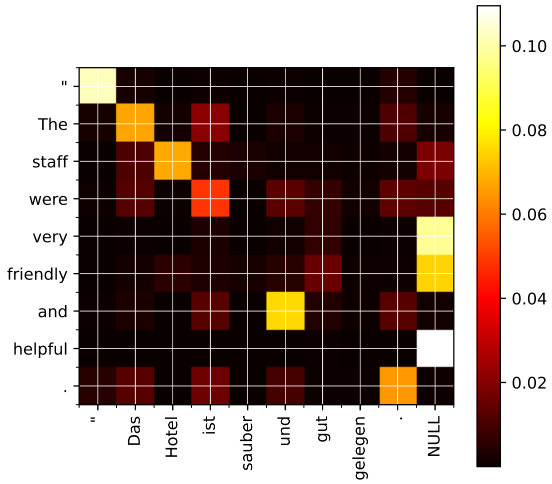

Figure 1: A hallucinatory German-English translation was detected by correctly identifying the null alignments with our Optimal Transport-based method. Here, if a target word is mapping to "NULL", it is likely to be hallucinated, as no source word supports its translation.

Existing methods for detecting hallucinations and omissions in MT either focus on analyzing the model's *internal* states (e.g. cross-attention) or rely on the output of *external* tools, such as cross-lingual similarity estimators (Feng et al., 2022; Heffernan et al., 2022) or MT quality estimators (Rei et al., 2020; Duquenne et al., 2023). On one hand, *internal* methods are limited to scenarios where white-box access to the MT system is assumed, and are further limited by dependencies on specific model architectures (e.g. attention-based). On the other hand, *external* methods heavily rely on scalar outputs from models optimized for related tasks, such as MT quality or sentence similarity, but not directly optimized for hallucination detection. While hallucinations and omissions constitute low-quality translations, the opposite does not always hold. Furthermore, despite their effectiveness in detecting hallucinations and omissions, existing *internal* and *external* methods cannot distinguish

*Equal contribution

†Work is done during an internship at Noah's Ark Lab.

[1]Our code is publicly available at https://github.com/chenyangh/OTTAWA

Findings of the Association for Computational Linguistics ACL 2024, pages 6322–6334
August 11-16, 2024

between these two error types. This is because they frame the task as anomaly detection, relying on a single *outlier score* to detect both error types.

Cross-lingual word alignment (Brown et al., 1993a; Och and Ney, 2003) aligns source to target tokens in parallel sentence pairs, and it provides unique insight for translation errors like hallucination and omission. Following the definition by Guerreiro et al. (2023c), hallucinated translation happens when the generated content is largely detached from the source text; therefore, we can measure hallucination by the amount of "misaligned" words in the translated text. Similarly, an omission error is likely if a significant amount of source words are not found in the translation.

Despite the strong relevance between word alignments and translation errors, alignment tools have not yet been widely applied for hallucination and omission detection. In particular, we observe that existing alignment approaches (Sabet et al., 2020; Azadi et al., 2023; Arase et al., 2023) are often based on the assumption that the translation is correct, such that most of the source or target words can be aligned with each other. This assumption, however, does not hold for scenarios where translation errors have occurred; hence, the direct application of existing word alignment tools does not work well for detecting translation errors.

In this paper, we utilize Optimal Transport problem (Villani, 2009; Cuturi, 2013) and propose OTTAWA, an Optimal TransporT Adaptive Word Aligner that is specifically tailored to aligning words for the pathological cases where significant hallucination and omission errors are present. Our approach explicitly models the missing alignments by introducing a "null" vector. Specifically, we propose a novel one-side constrained Optimal Transport (OT) problem, which allows the null vector to be adaptively aligned to any number of words. Moreover, we set the null vector as a centric point, which is not biased during the computation of the OT problem. We show an example of our OT-based method in Figure 1.

We conduct extensive experiments using the newly proposed HalOmi benchmark (Dale et al., 2023b), designed for hallucination and omission detection, across a total of 18 language pairs. Results show that our word aligner-based approach, equipped with OTTAWA, performs on par with state-of-the-art *internal* and *external* detection methods (Guerreiro et al., 2023b; Dale et al., 2023a). In contrast to previous approaches, we show that our method can distinguish between hallucination and omission errors, while OTTAWA being crucial for performances compared to using existing word aligners.

2 Related Work

Hallucination remains a persistent concern in machine translation systems, which have recently gained increased attention in the field. Zhou et al. (2021) studied token-level hallucinations in machine translation on a Chinese-English dataset, which they constructed by randomly corrupting some tokens in the translation sentences. Guerreiro et al. (2023c) developed a dataset of 3.4k naturally occurring German-English translations, manually annotated at both the sentence and token levels for hallucinations and omissions detection. HalOmi (Dale et al., 2023b) expanded the language coverage for this task by manually annotating hundreds of translation pairs across 18 language pairs.

Methods for detecting hallucinations and omissions in machine translation can be broadly categorized into two types: *internal*, which analyzes the translation model's own outputs and states, and *external*, which relies on additional tools or data outside the model itself (Dale et al., 2023a). Guerreiro et al. (2023b) developed a method that employs various optimal transport-based distances (Kantorovich, 1942; Villani, 2009) to evaluate the abnormality in a machine translation (MT) model's internal cross-attention distribution of a given translation, by making comparisons with those from a collection of high-quality translations. This method not only requires access to the internal states of the MT model but also necessitates a large collection of cross-attention maps from ground truth parallel sequences for the targeted language pair.

Dale et al. (2023a) investigated the use of other internal MT model components, such as length-normalized sequence log-probability (Seq-Logprob) and the token-token layer-wise interaction framework (ALTI) [2], as alternative abnormality estimators that do not require external translation data. In addition, the authors suggested directly using scores from *external* MT quality estimator models, such as COMET-QE (Rei et al., 2020), or employing the cosine similarity between the source and translation as generated by cross-lingual sentence similarity models, including LASER (Heffer-

[2]Originally proposed by (Ferrando et al., 2022) for interpreting MT systems

nan et al., 2022), LaBSE (Feng et al., 2022).

Word alignment (Brown et al., 1993a,b) has been extensively studied over the years in both cross-lingual (Nagata et al., 2020; Dou and Neubig, 2021) and monolingual (Nagata et al., 2020; Lan et al., 2021) settings. The task is approached with supervised learning methods (Och and Ney, 2003; Östling and Tiedemann, 2016), which require sentence-level parallel data, as well as unsupervised methods that do not rely on such data. The latter approach involves leveraging token-token attention matrices (Ghader and Monz, 2017; Zenkel et al., 2020; Zhang and van Genabith, 2021) or pairwise similarity matrices between tokens' embedding vectors (Sabet et al., 2020; Azadi et al., 2023), both extracted from a multilingual pretrained language models (Devlin et al., 2019; Conneau et al., 2020). The application of Optimal Transport (OT, Monge, 1781; Kantorovich, 1942; Villani, 2009; Cuturi, 2013) in natural language processing tasks was mainly focused on similarity estimation between textual segments (Huang et al., 2016; Alqahtani et al., 2021; Lee et al., 2022; Mysore et al., 2022), and more recently for word alignment (Azadi et al., 2023; Arase et al., 2023).

3 Method

We formulate the unsupervised word alignment task in § 3.1, and then provide an overview on how standard and partial optimal transport are employed to address the task in § 3.2 and § 3.3, respectively. In § 3.4, we describe our newly proposed optimal transport word aligner, OTTAWA, specifically adapted to enhance the detection of MT hallucinations and omissions.

3.1 Unsupervised Word Alignment

Given a source word sequence $\boldsymbol{s}$=$[s_1, \ldots, s_m]$ and a target word sequence $\boldsymbol{t}$=$[t_1, \ldots, t_n]$, let $E^{(s)}$=$[\boldsymbol{e}_1^{(s)}, \ldots, \boldsymbol{e}_m^{(s)}] \in \mathbb{R}^{m\times D}$ and $E^{(t)}$=$[\boldsymbol{e}_1^{(t)}, \ldots, \boldsymbol{e}_n^{(t)}] \in \mathbb{R}^{n\times D}$ be the embedding matrices for the source and target sequences, respectively, where D is the dimensionality of the embeddings. A cost matrix $C \in \mathbb{R}^{m\times n}$ is defined such that $C_{i,j}$ is the cosine distance between the i-th source vector and the j-th target vector.

The goal is to compute the alignment matrix $\Gamma \in \{0,1\}^{m\times n}$, where $\Gamma_{i,j} = 1$ if the i-th source word is aligned to the j-th target word, such that the aligned words have the smallest distance in the cost matrix C. A straightforward approach to obtain Γ involves making greedy decisions on each word alignment in both forward (source-to-target) and reverse (target-to-source) directions. Specifically, given the cost matrix C, the forward alignment $\Gamma^{(F)}_{i,j}$ can be obtained by:

$$\Gamma^{(F)}_{i,j} = \begin{cases} 1, & \text{if } j = \operatorname{argmin}_j C_{i,j} \\ 0, & \text{otherwise} \end{cases} \tag{1}$$

and similarly, the reverse alignment $\Gamma^{(R)}_{i,j}$ is computed by:

$$\Gamma^{(R)}_{i,j} = \begin{cases} 1, & \text{if } i = \operatorname{argmin}_i C_{i,j} \\ 0, & \text{otherwise} \end{cases} \tag{2}$$

The final alignment matrix Γ is obtained by taking the element-wise multiplication of $\Gamma^{(F)}$ and $\Gamma^{(R)}$. However, Sabet et al. (2020) found that greedy decisions tend to ignore the word pairs of relatively lower similarity. Therefore, the authors propose to formulate the word alignment as an assignment problem (Kuhn, 1955), which consequently involves solving the following optimization problem:

$$\Gamma^* = \operatorname*{arg\,min}_{\Gamma \in U^{(A)}} \sum_{i,j} C_{i,j}\Gamma_{i,j} \tag{3}$$

where $U^{(A)} = \{\Gamma \in \{0,1\}^{m\times n} : \Gamma \mathbf{1}_n \leq \mathbf{1}_n, \Gamma^\top \mathbf{1}_m \leq \mathbf{1}_m, \sum_{i,j}\Gamma_{i,j} = \min(m,n)\}$ is the set of all binary matrices with the sum of row and column equal to 1. The optimization in Eq. 3 can be solved by linear programming (Bourgeois and Lassalle, 1971).

3.2 Standard Optimal Transport

Chi et al. (2021) have noticed that binary assignments obtained by Eq. 3 pose challenges when aligning source and target texts of markedly different lengths. This is because the binary assignment assumes that each word in the shorter text corresponds to a word in the longer text, while the additional words in the longer text remain unaligned. To address this issue, Chi et al. (2021); Dou and Neubig (2021) leverage standard optimal transport as an alternative solution for word alignment. The goal is to compute a matrix $P^* \in \mathbb{R}_+^{m\times n}$ as follows:

$$P^* = \operatorname*{arg\,min}_{P \in U^{(O)}} \sum_{i,j} C_{i,j}P_{i,j} \tag{4}$$

where $U^{(O)} = \{P \in \mathbb{R}_+^{m\times n} : P\mathbf{1}_n = \boldsymbol{\mu}, P^\top \mathbf{1}_m = \boldsymbol{\nu}\}$, $\boldsymbol{\mu} = (1/m, \ldots, 1/m)$, and

$\boldsymbol{\nu} = (1/n, \dots, 1/n)$. The binary matrix Γ can be obtained by replacing the cost matrix C with the matrix P^* in Eq. 1 and Eq. 2, followed by performing element-wise multiplication on the resultant matrices. It is important to note that standard OT strictly requires the use of marginal constraints on $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$, which enforce the alignment of each source or target word. Therefore, computing forward and reverse alignments is crucial for performance, as these computations implicitly model null alignments. This necessity arises because standard OT-based word alignment assumes that there is a one-to-one mapping between each word in the source sequence and a corresponding word in the target sequence, and vice versa.

3.3 Partial Optimal Transport

The one-to-one mapping assumption of standard OT is not always applicable, particularly in tasks with unbalanced source-target word alignments, such as monolingual text summarization. In response, Arase et al. (2023) introduce partial optimal transport (Peyré et al., 2016; Chapel et al., 2020) word aligner that relaxes the one-to-one mapping assumption, explicitly allowing for source or target words to remain naturally unaligned. In their setting, the transport map P can be found by solving the optimization problem as in Eq. 4 but with different constraints:

$$U^{(P)} = \{P \in \mathbb{R}_{+}^{m \times n} : P\mathbf{1}_n \leq \boldsymbol{\mu}, \\ P^\top \mathbf{1}_m \leq \boldsymbol{\nu}, \sum_{i,j} P_{i,j} = m\} \tag{5}$$

where $\boldsymbol{\mu}$ and $\boldsymbol{\nu}$ are uniform distributions, $m \in (0, 1)$ is a hyperparameter controlling how many source and target words are aligned. Then, a threshold τ is used to obtain the binary alignment matrix Γ from the matrix P. Although the partial OT-based drops the strong assumption of the standard OT, it requires extensive search over the hyperparameters m and τ to determine whether a source and target word are considered aligned or not. Furthermore, the conditions in Eq. 5 favor the alignment of closely similar words, because the marginal constraints for either side are not guaranteed. Consequently, this promotes a high recall for alignments at the expense of precision when source and target words are not highly similar.

3.4 OTTAWA

Accurately identifying unaligned words in the target text is crucial for detecting MT hallucinations. Similarly, identifying unaligned words in the source text enhances the detection of omissions. We propose explicitly handle null alignments as a mapping to special *null* words, denoted as s_{m+1} and t_{n+1}, which are appended to the source $\mathbf{s}$ and target $\mathbf{t}$ sequences, respectively. We denote the embeddings corresponding to s_{m+1} and t_{n+1} as $\boldsymbol{e}_{m+1}^{(s)}$ and $\boldsymbol{e}_{n+1}^{(t)}$, respectively. Our approach addresses forward and reverse alignments separately, yet employs the same method for both. Therefore, we describe how $\Gamma^{(R)}$ is obtained, as $\Gamma^{(F)}$ can be derived analogously. It is important to note that in performing a reverse alignment, only $\boldsymbol{e}_{m+1}^{(s)}$ is appended to $E^{(s)}$, while $E^{(t)}$ remains unchanged, and vice versa for forward alignment.

Let $\boldsymbol{e}_{m+1}^{s}$ be denoted as $\boldsymbol{e}^{(\varnothing)}$ hereafter for simplicity. So far, $\boldsymbol{e}^{(\varnothing)}$ can be any vector but to make the OT problem meaningful, we need to impose constraints on the distance between $\boldsymbol{e}^{(\varnothing)}$ and the target word vectors $E^{(t)}$. Firstly, $\boldsymbol{e}^{(\varnothing)}$ should be equidistant to every target vector to avoid bias. In addition, these equal distances should be of the same scale as the average pair-wise distances between the source and target vectors to avoid aggressive alignment to $e^{(\varnothing)}$.

To this end, we extend the cost matrix C to $\bar{C} \in \mathbb{R}^{(m+1) \times n}$, where $\bar{C}_{i,j} = C_{i,j}$ for $i \in [1, m]$ and $\bar{C}_{m+1,j} = d$, where d is the distance between $\boldsymbol{e}^{(\varnothing)}$ and the target word representations (by our assumption, it is the same for all target words). One can define d in multiple ways. For example, a natural choice is to use the mean of pair-wise distances between the source and target vectors. However, we choose to use the median over the mean to avoid the influence of the outliers. It is also based on the intuition that $\boldsymbol{e}^{(\varnothing)}$ should serve as a *centric point*, which is not biased during the computation of the optimal transport map.

For $\boldsymbol{e}^{(\varnothing)}$ to be realizable in the vector space $\mathbb{R}^D$, the equal distance d has a lower bound d_{min} (see Appendix A). We finally define $d = \max(d_{\text{min}}, c)$, where c is the median of all pair-wise distances between the source and target words. We reformulate the optimization problem for the transport map in Eq. 4 by replacing the cost matrix C with $\bar{C}$ and relaxing the constraints as follows:

$$U^{(\varnothing,R)} = \{P \in \mathbb{R}_{+}^{(m+1) \times n} : P\mathbf{1}_n \leq \boldsymbol{\mu}', \\ P^\top \mathbf{1}_{m+1} = \boldsymbol{\nu}\} \tag{6}$$

where $\boldsymbol{\mu}' = (\boldsymbol{\mu}, 1)$ allows $\boldsymbol{e}^{(\varnothing)}$ to have a maximum marginal of 1, and to be aligned with any number of target vectors. We denote the solution of this optimization problem as $P^{(\varnothing,R)}$. The reverse alignment matrix $\Gamma^{(R)}$ is computed as in Eq. 2 by replacing C with $P^{(\varnothing,R)}$. The forward alignment matrix $\Gamma^{(F)}$ is computed analogously. The final alignment matrix Γ is obtained by taking the element-wise product between $\Gamma^{(F)}$ and $\Gamma^{(R)}$.

Our optimization problem in the set of Eq. 6 is a special case of partial optimal transport, it combines the properties of Eq. 4 and Eq. 5. For example, in the computation of $\Gamma^{(R)}$, we relax the marginal constraint on the source side, a necessity given that the marginal for $\boldsymbol{e}^{(\varnothing)}$ can range between 0 and 1. As a result, the model has the freedom to adaptively decide whether or not to make *null alignment* based on our defined median distance d. Conversely, the marginal constraint on the target side remains unchanged, ensuring that each target vector is compared with the centric point $\boldsymbol{e}^{(\varnothing)}$.

To apply word alignment for hallucination and omission detections, we propose to analyze the number of non-aligned words from the final alignment matrix Γ. For both source and target sentences, we obtain the ratio of the misaligned words, given by:

$$r^{(R)} = \frac{1}{m}\sum_{i=1}^{m} \mathbb{I}\left(\sum_{j=1}^{n} \Gamma_{i,j} > 0\right) \tag{7}$$

$$r^{(F)} = \frac{1}{n}\sum_{j=1}^{n} \mathbb{I}\left(\sum_{i=1}^{m} \Gamma_{i,j} > 0\right) \tag{8}$$

where $\mathbb{I}(\cdot)$ is the indicator function. Further, we accumulate the transported mass to $\boldsymbol{e}^{(\varnothing)}$ as the confidence of missing alignments, given by:

$$c^{(R)} = \frac{1}{n}\sum_{j=1}^{n} P^{(\varnothing,R)}_{m+1,j}\ ;\quad c^{(F)} = \frac{1}{m}\sum_{i=1}^{m} P^{(\varnothing,F)}_{i,n+1}$$

The combined score $r^{(R)} + c^{(R)}$ is used to detect hallucinations, and $r^{(F)} + c^{(F)}$ is for omissions.

4 Experimental Setup

4.1 Datasets and Evaluation Metrics

We conducted experiments using the recently proposed HalOmi benchmark (Dale et al., 2023b), which contains manually annotated data for MT hallucination and omission detection across 18 language pairs. It is constructed by pairing English sentences with translations in 5 high-resource languages (ar:Arabic, ru:Russian, es:Spanish, de:German, and zh:Mandarin) and 3 low-resource languages (ks:Kashmiri, mni:Manipuri, and yo:Yoruba), and a one zero-shot[3] Spanish-Yoruba pair. Each language pair includes between 145 to 197 examples, totaling 2,865 overall. Each example is annotated with 4 labels: *no*, *small*, *partial*, and *full* hallucination, with a similar scheme applied for omissions. Therefore, we utilized the multi-class ROC AUC variant, as defined in HalOmi (Dale et al., 2023b).

4.2 Baselines and Models

We conduct experiments to compare **internal** and **external** approaches for detecting hallucinations and omissions in MT, alongside our newly proposed **word aligner**-based approach. More precisely, consider Seq-Logprob, ALTI, ALTI$^{\mathrm{T}}$, and ATT-OT (originally proposed by (Guerreiro et al., 2023b)). Similarly, for external approaches, we report results from MT quality estimators LaBSE (Feng et al., 2022) and BLASER-QE (Barrault et al., 2023), as they are the best performers in this category according to (Dale et al., 2023b). We refer the readers to (Dale et al., 2023b) for a detailed of these baselines.

4.3 Implementation Details

Our implementation is based on PyTorch (Paszke et al., 2019), and we use the POT library[4] for all the Optimal Transport-based methods. For the entropy-regulized O (including the standard OT and partial OT), there is a regularization parameter ϵ that controls the sparsity of the OT solution, where 0.1 is the default value. In our experiments, we set ϵ to a low value of 0.05 to encourage the confidence of the sovler. However, we observed that the performance is not sensitive, but an overly small ϵ may lead to numerical instability.

In all experiments, we adopt a word-level representation approach which is consistent with standard practices in word alignment task (Sabet et al., 2020; Azadi et al., 2023). Specifically, we first generate token-level representations for each token in the sequence and establish a token-to-word index mapping. Then, we apply mean pooling to average the tokens representations corresponding to the same word, resulting in word-level representations. In our main experiment uses the representations of the last layer of LaBSE (Feng et al.,

[3]The MT system utilized for generating the data was not trained on Spanish-Yoruba parallel corpora.

[4]https://pythonot.github.io/

Source	Lang	*Hallucination*							*Omission*						
		$\mathcal{I}1$	$\mathcal{I}2$	$\mathcal{I}3$	$\mathcal{I}4$	$\mathcal{E}1$	$\mathcal{E}2$	**Our**	$\mathcal{I}1$	$\mathcal{I}2$	$\mathcal{I}3$	$\mathcal{I}4$	$\mathcal{E}1$	$\mathcal{E}2$	**Our**
High Resource	en-ar	89	78	54	36	84	90	**93**	63	44	76	64	79	**85**	79
	ar-en	83	75	72	51	88	**94**	90	56	49	82	72	**83**	77	70
	en-ru	**95**	36	37	61	86	89	91	53	56	79	75	**85**	84	**85**
	ru-en	86	76	56	55	83	92	**99**	59	52	86	76	76	80	**89**
	en-es	87	85	69	59	**88**	85	**88**	39	31	**89**	89	86	80	87
	es-en	92	89	67	37	**94**	87	87	61	50	59	66	**73**	71	**73**
	en-de	85	**97**	69	55	**97**	87	88	50	46	77	74	66	**85**	81
	de-en	90	80	59	65	95	**97**	96	48	38	82	**83**	70	70	74
	en-zh	**88**	82	47	60	86	78	**88**	60	51	73	67	69	88	**92**
	zh-en	**89**	88	65	46	88	87	86	73	61	**77**	64	75	73	**77**
Avg. High Resource		88	78	59	53	89	89	**91**	54	46	78	74	76	80	**81**
Low Resource	en-ks	68	71	74	54	76	**81**	78	50	52	**90**	81	76	77	84
	ks-en	59	67	65	65	57	**73**	56	36	50	64	63	59	45	**70**
	en-mni	68	81	49	54	80	**83**	59	45	52	80	73	**82**	80	69
	mni-en	70	64	49	56	56	**78**	64	42	33	68	63	52	74	**77**
	en-yo	77	74	59	65	56	79	**80**	77	50	70	72	65	67	**86**
	yo-en	78	72	54	43	66	**80**	65	**68**	60	65	55	66	56	**68**
Avg. Low Resouce		70	71	57	56	66	**79**	67	53	49	73	68	67	67	**76**
Zero-Shot	yo-es	60	**65**	47	44	56	57	54	62	47	**85**	69	62	66	76
	es-yo	61	66	52	55	66	**68**	65	68	50	**83**	60	69	67	80
Avg. Zero-Shot		60	**66**	49	49	61	63	59	65	48	**84**	65	66	67	78
Avg. HalOmi		79	75	57	53	78	**83**	79	56	48	77	70	72	74	**79**

Table 1: Methods performances (ROC AUC) on hallucination and omission detection across HalOmi's high-resource, low-resource, and zero-shot sets. Bold entries describe the best results among all methods, which we categorize under 3 approaches: $\mathcal{I}$nternal, $\mathcal{E}$xternal, and Our word alignment. $\mathcal{I}1, \mathcal{I}2, \mathcal{I}3, \mathcal{I}4$: Seq-Logprob, ALTI, ALTI$^{\mathrm{T}}$, and ATT-OT. $\mathcal{E}1, \mathcal{E}2$: LaBSE and BLASER-QE MT quality estimators. Our: OTTAWA using LaBSE token embeddings. All scores are scaled within the range of $[0, 100]$ following (Dale et al., 2023b).

2022). When experimenting with other models, we use the hidden representation of the 8th layer of mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2020), respectively, following (Dou and Neubig, 2021; Chi et al., 2021; Azadi et al., 2023).

Our OTTAWA, PMIAlign, and the standard OT do need extra hyper-parameters to determine word alignments. However, for partial OT, we set $m = 0.5$ and tested a few thresholds. Specifically, we set τ to 0.1, 0.05, and 0.025 of the maximal value in marginal μ and ν, given by $\max(1/m, 1/n)$, and used the best threshold 0.05 to report the scores. To obtain the hallucination and omission, all baseline aligners use the reverse missing alignment ratio Eq. 7 to estimate hallucination and use Eq. 8 to obtain omission scores.

5 Results

5.1 Sentence-Level Detection

Table 1 shows the performances of methods across 3 approaches for hallucination and omission detection across HalOmi's 18 language pairs: 4 methods representing the *internal* approach, 2 for the *external* approach, and 1 showcasing our newly introduced *word alignment* approach. The latter approach employs our OTTAWA word aligner, which leverages LaBSE to obtain token-level representations.

Although our method performs best on only 5 out of 18 language pairs for hallucination and 7 out of 18 for omission, it outperforms other methods in terms of overall average performance. More precisely, we outperform all baselines in high-resource languages for both hallucination and omission detection. However, we underperform in low-resource languages (except for omission detection) and zero-shot language pairs. Overall, although we underperform by 4% compared to BLASER-QE ($\mathcal{E}2$) on average hallucination detection across the 18 language pairs, we surpass that model by 5% on omission. It is worth mentioning that BLASER-QE is a strong baseline, it is an MT quality estimator that leverages the SONAR (Duquenne et al., 2023) sentence embeddings. It was carefully tuned on 123.4k source-translation pairs, derived from outputs of 24 MT systems and covering 62 language pairs, all manually labeled by humans for MT quality estimation.

However, our method significantly improves upon LaBSE ($\mathcal{E}2$), another MT quality estimator. While we report close tight (1% gain on HalOmi average) on hallucination, we significantly outperform it by 6% on omission. These results demonstrate that LaBSE's token-level representations contain more expressive information for detecting omissions compared to the scalar values generated for MT quality estimation.

Finally, we observe that our method provides consistent and balanced performance in detecting both hallucinations and omissions compared to *internal* methods. For instance, although Seq-Logprob ($\mathcal{I}1$) matches our method on hallucinations, it significantly underperforms compared to us by 23% on omissions. Similarly, ALTI$^{\mathrm{T}}$ ($\mathcal{I}3$) has the smallest gap (2%) with our method on omission but performs poorly, achieving only 57% compared to our method's 79%, on hallucination.

5.2 Ablation Study

We conduct ablation studies on our approach, testing variants that modify its two key components: the **Embedding** vector representation model and the word **Aligner**. Table 2 presents the average scores for both HalOmi high-resource (HR) and low-resource (LR) language pair clusters, along with the overall average. The complete results across the 18 language pairs are presented in Table 5 in Appendix B.

	Hallucination			*Omission*		
	HR	**LR**	**Avg**	**HR**	**LR**	**Avg**
Embedding						
mBERT	87	56	73	**81**	69	76
XLMR	88	53	71	78	64	72
LaBSE	**91**	**67**	**79**	**81**	**76**	**79**
Aligner						
SimAlign	47	55	51	70	62	66
PMIAlign	81	61	72	**82**	71	78
OT	81	60	72	**82**	74	**79**
POT	64	52	57	73	60	68
OTTAWA	**91**	**67**	**79**	81	**76**	**79**

Table 2: Hallucation and omission detection average AUC scores on HalOmi's high-resource (HR) and low-resource (LR) languages. **Embeddings** refer to the results obtained by testing OTTAWA with token embeddings from various models. **Aligner** are the results when employing different word aligners, all utilizing LaBSE embeddings.

Embedding We conduct experiments with OTTAWA, utilizing embeddings derived from unsupervised pre-trained models such as mBERT (Devlin et al., 2019) and XLM-R (Conneau et al., 2020). These models are commonly used in cross-lingual word alignment (Sabet et al., 2020; Azadi et al., 2023). We notice that OTTAWA with LaBSE representations significantly outperforms the best-unsupervised representation baseline (mBERT) by 6% and 3% on the overall HalOmi average score on hallucination and omission, respectively. This is expected given that LaBSE representations are more task-specific than those from unsupervised models, having been extensively fine-tuned for translation pair quality estimation using carefully curated data for low-resource languages.

Consequently, the performance gap between LaBSE and models such as mBERT is larger in low-resource languages (11% and %3 on omissions) compared to high-resource languages (4% and 0% on hallucinations). Overall, the results are promising, demonstrating that hallucinations and omissions can be detected even without access to the MT system or a robust MT quality estimator. This approach serves as an acceptable alternative, particularly in high-resource settings. However, its performance in low-resource languages necessitates further investigation.

Aligner We compared OTTAWA with state-of-the-art word aligner methods, including SimAlign (Sabet et al., 2020) (Itermax), PMI-

Align (Azadi et al., 2023), standard OT (Chi et al., 2021; Dou and Neubig, 2021), and POT (Arase et al., 2023), all utilizing LaBSE representations.

We first observe that the standard alignment methods work reasonably well on detecting omissions but not hallucinations. This is an understandable outcome, which can be explained as follows: 1) the hallucinated translations often contain a significant amount of "detached" words. For example, the word "cat" may be translated to "dog". This is particularly problematic for standard alignment methods as they assume the word "cat" should have a corresponding word in the translation, resulting in aligning "cat" to "dog" nevertheless. 2) the omission errors in the HalOmi dataset are often early terminated translations. In this case, the translated words are still mapped to the source words, and the standard methods are still able to align the existing words and detect the omission errors.

However, only our method works well in the detection of both hallucinations and omissions, where there is a 7% overall average improvement over both PMIAlign and standard OT in hallucination errors. This outcome is consistent with our motivation that the common assumption among standard word-alignment tools, which is that a mapping exists between all source and target words, prevents them from effectively identifying all errors.

5.3 Distinguish Hallucination and Omission

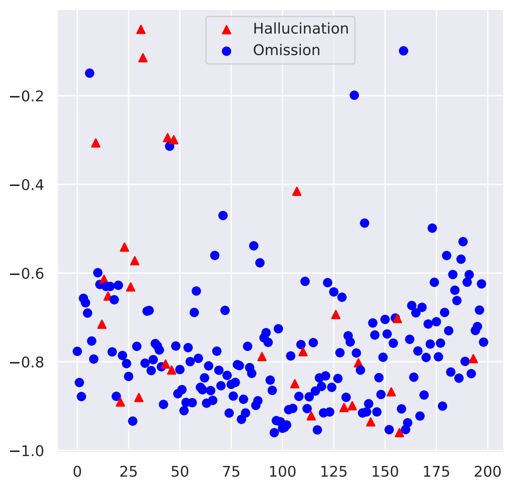

Figure 2: LaBSE cosine similarity score (y-axis) of samples, aggregated across 10 high-resource datasets, with either hallucination (red triangles) or omission (blue circles) as gold labels. The x-axis shows the sample count index.

Figure 2 shows LaBSE cosine similarity scores for samples with either hallucination or omission labels, aggregated from the 10 high-resource language pairs datasets. We only focus on high-resource languages, given that the performance of all methods significantly diminishes with low-resource ones. The primary goal of this experiment is to study the models' abilities to differentiate between types of errors. Although LaBSE quality estimator can effectively detect both hallucinations and omissions, the figure clearly shows that it fails to distinguish between them. The majority of samples for both hallucinations and omissions exhibit cosine similarity scores below -0.8.

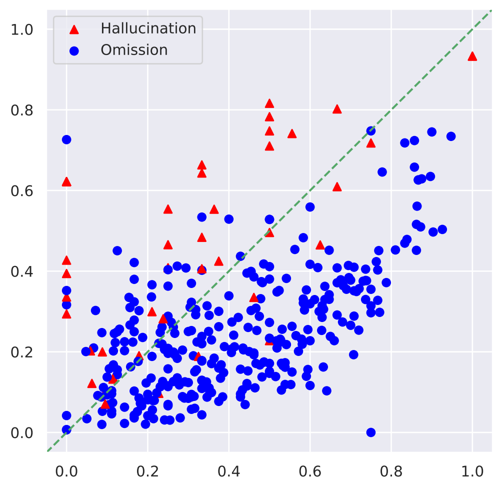

Figure 3: Hallucination scores (x-axis) and Omission scores (y-axis) produced by our OTTAWA across the same samples used in Figure 2. The dotted green line simply illustrates the diagonal.

Figure 3 shows the hallucination (x-axis) and omission (y-axis) scores produced by OTTAWA(using LaBSE embeddings), on the same samples used in Figure 2. The figure demonstrates that our approach effectively distinguishes between omission and hallucination errors, with omission errors concentrated below the diagonal, and hallucination errors concentrated above it. In most cases, our word alignment approach, which operates without any supervision signal, accurately distinguishes omissions from hallucinations, and vice versa. This unique feature offered by our approach can help researchers to precisely diagnose and comprehend error types in MT systems, thereby facilitating more efficient development and the implementation of targeted solutions.

5.4 Word-Level Detection

Following the experimental setup in (Dale et al., 2023b), we evaluate our OTTAWA for word-level hallucination and omission detection. Table 3 compares its performance with the best *internal* approach baselines identified in their study: LogProb (the `Stand.` variant) and ALTI (the `sum` variant). It is worth noting that the LogProb and ALTI baselines are different from those in Table 1 despite the name resemblance. Specifically, ALTI is a combination of two baselines from Table 1: ALTI (hallucination results), and ALTI^{T} (omission results). Conversely, the LogProb baseline is obtained by processing the sentence pairs through the MT model twice (once for the source and once for the target), unlike SeqLogProb in Table 1, which involves only one pass.

	Hallucination			*Omission*		
	HR	**LR**	**Avg**	**HR**	**LR**	**Avg**
LogProb	73	66	70	84	**71**	78
ALTI	**87**	**69**	**78**	**86**	69	**78**
OTTAWA	78	59	70	79	60	71

Table 3: Word-level hallucination and omission results, evaluated using the word-level ROC AUC score as defined by (Dale et al., 2023b).

Additionally, (Dale et al., 2023b) conducted ablations with these baselines, creating two variants for each: Standard and Contrastive for LogProb, and Mean and Max for ALTI. In Table 3, we report only the results for the best-performing variant of each baseline. Results indicate that word-level detection poses a greater challenge for our approach, which significantly under-performs compared to the best baseline (ALTI) by 8% and 7% in the overall average for hallucination and omission detection, respectively. However, we find the results encouraging when considering that unsupervised world-level detection with current state-of-the-art methods was not feasible without white-box access to the MT system, as *external* approaches operate on sentence level only.

5.5 Word Alignment

We run experiments on cross-lingual word alignment using 6 datasets, as in SimAlign (Sabet et al., 2020) and PMIAlign (Azadi et al., 2023), with gold word alignment labels. We includes English-Czech (en-cs) (Mareček, 2011), German-English (de-en) (Koehn, 2005), English-Persian (en-fa) (Tavakoli and Faili, 2014), English-French (en-fr) (Och and Ney, 2000), English-Hindi (en-hi) (Koehn et al., 2005) and Romanian-English (ro-en) (Mihalcea and Pedersen, 2003) language pairs. For evaluation metrics, we utilize the standard Alignment Error Rate (AER) (Och and Ney, 2003), and aligners use the mBERT representations.

	en-cs	**de-en**	**en-fa**	**en-fr**	**en-hi**	**ro-en**
	Alignment Error Rate (AER) ↓					
SimAlign	**0.12**	0.19	0.37	0.06	0.44	0.34
PMIAlign	**0.12**	**0.17**	**0.32**	0.06	**0.39**	0.31
OT	**0.12**	**0.17**	**0.32**	0.06	**0.39**	0.39
POT	0.13	0.19	0.37	0.06	0.44	0.34
OTTAWA	**0.12**	**0.17**	0.33	**0.05**	**0.39**	**0.30**

Table 4: Cross-lingual word alignment performances (AER) across 6 language pairs for 5-word aligners.

As shown in Table 4, OTTAWA performs on par with another state-of-the-art word aligner on the cross-lingual alignment task. More precisely, it matches the best-performing baselines in three languages (en-cs, de-en, and en-hi), while slightly outperforming them by 1% in the en-fr and ro-en language pairs. We anticipated no significant gains in word alignment, as the datasets for cross-lingual word alignment are constructed from clean translations, making null alignments less concerning.

6 Conclusion and Future Work

In this work, we propose a new word alignment-based approach for detecting hallucinations and omissions in MT systems. We develop OTTAWA, an innovative word aligner designed specifically for this purpose. While experiments show promise in MT hallucination and omission detection, this area remains an intriguing direction for future research exploration. We plan to focus on one-to-many alignments for pathological translations.

7 Limitations

Our Optimal Transport (OT)-based word alignment method relies on pretrained word embeddings, which may not be available for low and scarce resource languages. Also, the performance of our word aligner on low resources heavily depends on representations obtained from a supervised MT quality estimator, which may not be accessible or does not support certain languages. Another limitation of our study is the exclusive focus on null alignments, which are central to our task of interest. However, we do not address other complex cases such as one-to-many and many-to-many alignments.

Acknowledgements

We thank the anonymous reviewers for their insightful comments. Osmar R. Zaïane is supported by the Amii Fellow Program.

References

Sawsan Alqahtani, Garima Lalwani, Yi Zhang, Salvatore Romeo, and Saab Mansour. 2021. Using optimal transport as alignment objective for fine-tuning multilingual contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 3904–3919.

Yuki Arase, Han Bao, and Sho Yokoi. 2023. Unbalanced optimal transport for unbalanced word alignment. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3966–3986.

Fatemeh Azadi, Heshaam Faili, and Mohammad Javad Dousti. 2023. Pmi-align: Word alignment with pointwise mutual information without requiring parallel training data. In *Findings of the Association for Computational Linguistics: ACL 2023*, pages 12366–12377.

Loïc Barrault, Yu-An Chung, Mariano Cora Meglioli, David Dale, Ning Dong, Paul-Ambroise Duquenne, Hady Elsahar, Hongyu Gong, Kevin Heffernan, John Hoffman, et al. 2023. Seamlessm4t-massively multilingual & multimodal machine translation. *ArXiv Preprint*.

F. Bourgeois and J. Lassalle. 1971. An extension of the munkres algorithm for the assignment problem to rectangular matrices. *Communications of the ACM*, 14(12):802–804.

Peter E Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993a. The mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*, 19(2).

Peter F Brown, Vincent J Della Pietra, Stephen A Della Pietra, and Robert L Mercer. 1993b. Advances in probabilistic and other parsing technologies. *IBM Journal of Research and Development*, 37(3):288–307.

Stephen L. Campbell and Carl Dean Meyer. 2009. *Generalized inverses of linear transformations*. Society for Industrial and Applied Mathematics.

Laetitia Chapel, Mokhtar Z. Alaya, and Gilles Gasso. 2020. Partial optimal transport with applications on positive-unlabeled learning. In *Advances in Neural Information Processing Systems*, volume 33, pages 9778–9789.

Zewen Chi, Li Dong, Bo Zheng, Shaohan Huang, Xian-Ling Mao, He-Yan Huang, and Furu Wei. 2021. Improving pretrained cross-lingual language models via self-labeled word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 3418–3430.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Édouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In *Advances in Neural Information Processing Systems*, volume 26.

David Dale, Elena Voita, Loic Barrault, and Marta R. Costa-jussà. 2023a. Detecting and mitigating hallucinations in machine translation: Model internal workings alone do well, sentence similarity Even better. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 36–50.

David Dale, Elena Voita, Janice Lam, Prangthip Hansanti, Christophe Ropers, Elahe Kalbassi, Cynthia Gao, Loic Barrault, and Marta Costa-jussà. 2023b. HalOmi: A manually annotated benchmark for multilingual hallucination and omission detection in machine translation. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing*, pages 638–653.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Zi-Yi Dou and Graham Neubig. 2021. Word alignment by fine-tuning embeddings on parallel corpora. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 2112–2128.

Paul-Ambroise Duquenne, Holger Schwenk, and Benoit Sagot. 2023. Sonar: Sentence-level multimodal and language-agnostic representations. *ArXiv Preprint*.

Fangxiaoyu Feng, Yinfei Yang, Daniel Cer, Naveen Arivazhagan, and Wei Wang. 2022. Language-agnostic bert sentence embedding. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 878–891.

Javier Ferrando, Gerard I Gállego, Belen Alastruey, Carlos Escolano, and Marta R Costa-jussà. 2022. Towards opening the black box of neural machine translation: Source and target interpretations of the transformer. In *Proceedings of the 2022 Conference on*

Empirical Methods in Natural Language Processing, pages 8756–8769.

Hamidreza Ghader and Christof Monz. 2017. What does attention in neural machine translation pay attention to? In *Proceedings of the Eighth International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 30–39.

Nuno M Guerreiro, Duarte M Alves, Jonas Waldendorf, Barry Haddow, Alexandra Birch, Pierre Colombo, and André FT Martins. 2023a. Hallucinations in large multilingual translation models. *Transactions of the Association for Computational Linguistics*, 11:1500–1517.

Nuno M. Guerreiro, Pierre Colombo, Pablo Piantanida, and André Martins. 2023b. Optimal transport for unsupervised hallucination detection in neural machine translation. In *Proceedings of the 61st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 13766–13784.

Nuno M Guerreiro, Elena Voita, and André FT Martins. 2023c. Looking for a needle in a haystack: A comprehensive study of hallucinations in neural machine translation. In *Proceedings of the 17th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1059–1075.

Kevin Heffernan, Onur Çelebi, and Holger Schwenk. 2022. Bitext mining using distilled sentence representations for low-resource languages. In *Findings of the Association for Computational Linguistics: EMNLP 2022*, pages 2101–2112.

Gao Huang, Chuan Guo, Matt J Kusner, Yu Sun, Fei Sha, and Kilian Q Weinberger. 2016. Supervised word mover's distance. *Advances in Neural Information Processing Systems*, 29.

L. V. Kantorovich. 1942. On the translocation of masses. *Doklady Akademii Nauk*, 37(2):227–229.

Philipp Koehn. 2005. Europarl: A parallel corpus for statistical machine translation. In *Proceedings of Machine Translation Summit X: Papers*, pages 79–86.

Philipp Koehn, Joel Martin, Rada Mihalcea, Christof Monz, and Ted Pedersen. 2005. Proceedings of the acl workshop on building and using parallel texts. In *Proceedings of the ACL Workshop on Building and Using Parallel Texts*.

Harold W Kuhn. 1955. The hungarian method for the assignment problem. *Naval Research Logistics Quarterly*, 2(1-2):83–97.

Wuwei Lan, Chao Jiang, and Wei Xu. 2021. Neural semi-markov crf for monolingual word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 6815–6828.

Seonghyeon Lee, Dongha Lee, Seongbo Jang, and Hwanjo Yu. 2022. Toward interpretable semantic textual similarity via optimal transport-based contrastive sentence learning. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 5969–5979.

David Mareček. 2011. Automatic alignment of tectogrammatical trees from czech-english parallel corpus. *Univerzita Karlova, Matematicko-fyzikální fakulta*.

Rada Mihalcea and Ted Pedersen. 2003. An evaluation exercise for word alignment. In *Proceedings of the HLT-NAACL 2003 Workshop on Building and using parallel texts: data driven machine translation and beyond*, pages 1–10.

Gaspard Monge. 1781. Mémoire sur la théorie des déblais et des remblais. *Histoire de l'Académie Royale des Sciences de Paris*.

Mathias Müller and Rico Sennrich. 2021. Understanding the properties of minimum bayes risk decoding in neural machine translation. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 259–272.

Sheshera Mysore, Arman Cohan, and Tom Hope. 2022. Multi-vector models with textual guidance for fine-grained scientific document similarity. In *Proceedings of the 2022 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4453–4470.

Masaaki Nagata, Katsuki Chousa, and Masaaki Nishino. 2020. A supervised word alignment method based on cross-language span prediction using multilingual bert. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 555–565.

Franz Josef Och and Hermann Ney. 2000. Improved statistical alignment models. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 440–447.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational linguistics*, 29(1):19–51.

Robert Östling and Jörg Tiedemann. 2016. Efficient word alignment with markov chain monte carlo. *The Prague Bulletin of Mathematical Linguistics*.

Adam Paszke, Sam Gross, Francisco Massa, Adam Lerer, James Bradbury, Gregory Chanan, Trevor Killeen, Zeming Lin, Natalia Gimelshein, Luca Antiga, et al. 2019. PyTorch: An imperative style, high-performance deep learning library. *Advances in Neural Information Processing Systems*, 32.

Gabriel Peyré, Marco Cuturi, and Justin Solomon. 2016. Gromov-wasserstein averaging of kernel and distance matrices. In *International Conference on Machine Learning*, pages 2664–2672.

Vikas Raunak, Arul Menezes, and Marcin Junczys-Dowmunt. 2021. The curious case of hallucinations in neural machine translation. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 1172–1183.

Ricardo Rei, Craig Stewart, Ana C Farinha, and Alon Lavie. 2020. COMET: A neural framework for mt evaluation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 2685–2702.

Masoud Jalili Sabet, Philipp Dufter, François Yvon, and Hinrich Schütze. 2020. Simalign: High quality word alignments without parallel training data using static and contextualized embeddings. In *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 1627–1643.

Leila Tavakoli and Heshaam Faili. 2014. Phrase alignments in parallel corpus using bootstrapping approach. *International Journal of Information and Communication Technology Research*.

Cédric Villani. 2009. *Optimal transport: Old and new*, volume 338. Springer.

Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms giza++. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1605–1617.

Jingyi Zhang and Josef van Genabith. 2021. A bidirectional transformer based alignment model for unsupervised word alignment. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 283–292.

Chunting Zhou, Graham Neubig, Jiatao Gu, Mona Diab, Francisco Guzmán, Luke Zettlemoyer, and Marjan Ghazvininejad. 2021. Detecting hallucinated content in conditional neural sequence generation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 1393–1404.

A Equidistant Vector

Assume that we have vectors $\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n$ in $\mathbb{R}^D$. Here we discuss how to find an equidistant vector $\boldsymbol{e}^{(\varnothing)}$ with respect to cosine distance. For that we need to solve a system of $N-1$ equations:

$$\textbf{dist}(e^{\varnothing}, \boldsymbol{e}_1) = \textbf{dist}(e^{\varnothing}, \boldsymbol{e}_j)$$

for all $j = 2, \ldots, N$.

Recall the definition of the cosine distance:

$$\textbf{dist}(\boldsymbol{x}, \boldsymbol{y}) = 1 - \frac{\boldsymbol{x} \cdot \boldsymbol{y}}{\|\boldsymbol{x}\| \|\boldsymbol{y}\|}.$$

Hence, the system of equations we need to solve is equivalent to the system of linear equations:

$$\frac{\boldsymbol{e}^{\varnothing} \cdot \boldsymbol{e}_1}{\|\boldsymbol{e}^{\varnothing}\| \|\boldsymbol{e}_1\|} = \frac{\boldsymbol{e}^{\varnothing} \cdot \boldsymbol{e}_j}{\|\boldsymbol{e}^{\varnothing}\| \|\boldsymbol{e}_j\|}.$$

Slightly rewriting it, we get:

$$\boldsymbol{e}^{\varnothing} \cdot \left(\frac{\boldsymbol{e}_1}{\|\boldsymbol{e}_1\|} - \frac{\boldsymbol{e}_j}{\|\boldsymbol{e}_j\|} \right) = 0.$$

Note that usually for our word alignment problems $N < D$, hence this system always has a solution. However, we want to find the minimum possible equal distance so that the NULL vector is meaningful for the OT task. For that, assume that $\boldsymbol{e}^{(\varnothing)}$ lies in the span of the given vectors: $\boldsymbol{e}^{\varnothing} = \sum_{k=1}^{N} a_k \boldsymbol{e}_k$. Inserting it in the previous system, we obtain a new homogeneous system on the coefficients a_k:

$$\boldsymbol{E}\boldsymbol{a} = \boldsymbol{0},$$

where $E_{jk} = \boldsymbol{e}_k \cdot \left(\frac{\boldsymbol{e}_1}{\|\boldsymbol{e}_1\|} - \frac{\boldsymbol{e}_j}{\|\boldsymbol{e}_j\|} \right)$ is a $(N-1) \times N$ matrix. For general positions of word representation vectors, the kernel of $\boldsymbol{E}$ is one-dimensional. Projection onto it could be computed as $\boldsymbol{I}_N - \boldsymbol{E}^{+}\boldsymbol{E}$, where $\boldsymbol{E}^{+}$ is Moore-Penrose inverse and $\boldsymbol{I}_N$ is $N \times N$ identity matrix (Campbell and Meyer, 2009). In this case $\boldsymbol{a}$ will be the basis vector for the kernel subspace. After finding $\boldsymbol{a}$, one can construct $e^{\varnothing} = \sum_{k=1}^{N} a_k \boldsymbol{e}_k$ and compute $d_{\text{min}} = \textbf{dist}(e^{\varnothing}, e_1)$. Note that there is no equidistant vector with a smaller distance than d_{min}.

B Additional Results

We present the comparisons of different word embeddings and alignment methods in Table 5.

Source	Lang	*Hallucination*							*Omission*						
		$\mathcal{W}1$	$\mathcal{W}2$	$\mathcal{W}3$	$\mathcal{W}4$	**Our**	$\mathcal{R}1$	$\mathcal{W}2$	$\mathcal{W}1$	$\mathcal{W}2$	$\mathcal{W}3$	$\mathcal{W}4$	**Our**	$\mathcal{R}1$	$\mathcal{W}2$
High Resource	en-ar	33	64	66	65	**93**	85	88	56	80	82	78	79	**84**	81
	ar-en	49	83	82	69	**90**	85	88	71	**74**	67	65	70	68	63
	en-ru	31	74	75	71	91	87	**94**	74	83	82	76	**85**	83	76
	ru-en	51	81	80	50	**99**	97	91	75	87	**90**	80	89	85	81
	en-es	54	88	86	77	**88**	85	84	64	88	**89**	70	87	88	89
	es-en	60	83	**88**	67	87	83	83	69	**79**	75	65	73	73	68
	en-de	40	86	86	52	**88**	81	83	65	78	81	78	81	82	**84**
	de-en	62	94	95	67	**96**	92	94	75	79	**80**	64	74	75	75
	en-zh	39	72	72	59	**88**	87	85	84	92	**94**	79	92	93	88
	zh-en	63	85	83	53	86	**87**	77	61	**80**	77	72	77	79	66
Avg. High Resouce		47	81	81	64	**91**	87	88	70	**82**	82	73	81	81	78
Low Resource	en-ks	51	62	65	47	**78**	62	52	62	84	78	46	**84**	83	71
	ks-en	51	53	51	44	**56**	56	56	56	66	**71**	49	70	61	68
	en-mni	55	56	56	51	**59**	50	52	55	63	67	65	**69**	60	49
	mni-en	57	59	56	56	**64**	50	52	66	67	74	61	**77**	65	63
	en-yo	55	77	74	60	**80**	67	57	69	82	85	75	**86**	76	72
	yo-en	57	59	59	52	**65**	60	45	61	66	68	62	68	**69**	59
Avg. Low Resouce		55	61	60	52	**67**	56	53	62	71	74	60	**76**	69	64
Zero-Shot	yo-es	51	54	54	47	**54**	48	45	72	76	77	**80**	76	76	73
	es-yo	62	62	63	56	**65**	57	56	59	80	**85**	53	80	76	72
Avg. Zero-Shot		56	57	57	52	**59**	53	51	66	78	**81**	67	78	76	72
Avg. HalOmi		51	72	72	57	**79**	73	71	66	78	**79**	68	**79**	76	72

Table 5: Methods performances (ROC AUC) on hallucination and omission detection across HalOmi's high-resource, low-resource, and zero-shot language-pair clusters. Bold entries describe the best results among all methods. OTTAWA refer to using OTTAWA word aligner with LABSE embeddings. $\mathcal{W}1, \mathcal{W}2, \mathcal{W}3, \mathcal{W}4$: SimAlign-Itermax (Sabet et al., 2020), PMIAlign (Azadi et al., 2023), OT (Azadi et al., 2023), POT (Arase et al., 2023) word aligners with LaBSE embeddings. $\mathcal{R}1, \mathcal{R}2$: OTTAWA with embeddings obtained from mBERT (Devlin et al., 2019) and XLMR (Conneau et al., 2020).