

Alignment via Mutual Information

Shinjini Ghosh^α, Yoon Kim^α, Ramón Fernández Astudillo^β, Tahira Naseem^β, Jacob Andreas^α
^αMIT ^βIBM

{shinghos,yoonkim}@mit.edu, {ramon.astudillo,tnaseem}@ibm.com, jda@mit.edu

Abstract

Many language learning tasks require learners to infer correspondences between data in two modalities. Often, these alignments are many-to-many and context-sensitive. For example, translating into morphologically rich languages requires learning not just how words, but morphemes, should be translated; words and morphemes may have different meanings (or groundings) depending on the context in which they are used. We describe an information-theoretic approach to context-sensitive, many-to-many alignment. Our approach first trains a masked sequence model to place distributions over missing spans in (source, target) sequences. Next, it uses this model to compute pointwise mutual information between source and target spans conditional on context. Finally, it aligns spans with high mutual information. We apply this approach to two learning problems: character-based word translation (using alignments for joint morphological segmentation and lexicon learning) and visually grounded reference resolution (using alignments to jointly localize referents and learn word meanings). In both cases, our proposed approach outperforms both structured and neural baselines, showing that conditional mutual information offers an effective framework for formalizing alignment problems in general domains.

1 Introduction

Natural language is compositional: meanings of complex utterances can be constructed by combining the meanings of their atomic constituents (Montague, 1973). As a consequence, many canonical language learning problems, from machine translation to grounded word learning, require learners to infer what these constituents are, and how they **align** across modalities (e.g. between English and Spanish, or English and the visual world).

Fig. 1 shows an example: in order to translate *discography* into Spanish, it is necessary to know

that the morpheme *graph* should be translated into *graf*, the affix *y* into *ía*, etc. Formally, given paired data (\mathbf{x}, \mathbf{y}) (e.g. sentences and translations) an alignment algorithm must return a collection of span pairs $\{(\mathbf{x}_i, \mathbf{y}_i)\}$ where each \mathbf{x}_i and \mathbf{y}_i are contiguous sub-sequences of \mathbf{x} and \mathbf{y} respectively, and have the same meaning.

Most alignment algorithms assume that both \mathbf{x} and \mathbf{y} are sequences pre-segmented into words or word pieces (e.g. Brown et al., 1990; Zenkel et al., 2020), and that phrase-level alignments are ultimately reducible to word-level ones. But this assumption is quite restrictive: it limits these algorithms’ applicability in languages with complex morphology or where segmentation is otherwise more complex. More importantly, it means that these algorithms cannot be applied to problems involving non-linguistic (e.g. visual) data, in which it is possible that every observed fragment of an input will consist of a unique observation (e.g. set of pixel values). Indeed, we are not aware of any existing alignment algorithms that can be applied agnostically in both settings. Many alignment also make strong context-independence assumptions—for example, that each word in a sentence is translated or interpreted independently. This assumption can make it difficult to infer alignments in problems where language use is highly contextual (e.g. in the presence of polysemy, Thompson et al., 2018; or pragmatic constraints, Hickey, 1998).

How might we formulate the alignment problem in a way that accommodates unknown segment boundaries, context-dependence, and both linguistic and non-linguistic data? In this paper, we offer an information-theoretic framing of alignment: segments \mathbf{x}_i and \mathbf{y}_i are aligned if they have high pointwise mutual information (PMI) in the contexts where they occur. This approach avoids assumptions about data modality and segmentation (as PMI is straightforward to calculate for arbitrary spans in inputs of arbitrary types), and about con-

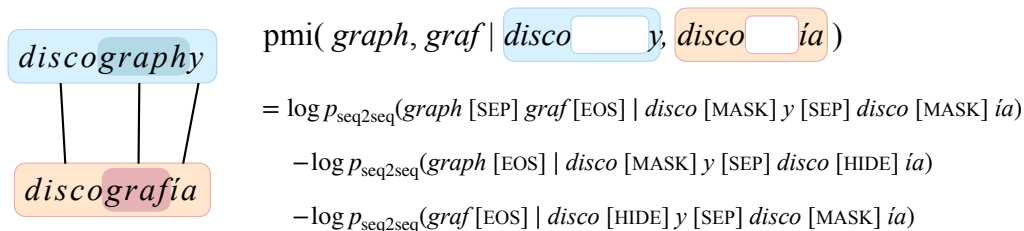


Figure 1: INFOALIGN: deriving alignments from conditional pointwise mutual information (PMI) using masked language models. We first train neural sequence models to reconstruct masked portions of paired sequences (e.g. characters forming words). This model is trained to assign probabilities to pairs of masked regions jointly and marginally (as in the two bottom terms on the right side of the figure). Finally, these models are used to compute conditional PMI between arbitrary span pairs. We use these scores to extract bilingual lexicons and resolve references in grounded tasks.

ditional independence (as PMI can be computed conditional on a linguistic or perceptual context).

Our approach, which we call INFOALIGN, first trains masked sequence models to compute joint and marginal probabilities of sub-spans of x and y in context, then uses these models to compute mutual information between spans. Using this span-scoring procedure, we define algorithms for extracting flat or hierarchical correspondences between modalities. We use these extracted alignments for two tasks: learning a morpheme-level lexicon to support zero-shot word translation in a low-data setting, and learning grounded representations of word meaning in a pragmatic reference task. In both settings, INFOALIGN outperforms both neural and structured approaches to learning many-to-many alignments.

2 Background and related work

At an intuitive level, given joint distribution over pairs of sequences (x, y) , alignment algorithms seek to find correspondences between “pieces” of x and y . Depending on the nature of the task, the granularity of these pieces may or may not be known. For instance, word or character level alignments operate over well-defined units. Morpheme or phrase alignments, on the other hand, often require joint induction of alignments and the units themselves.

Generative alignment models Some of the earliest alignment models came from the machine translation literature (e.g. Brown et al., 1990), which define generative models of sentences in a source language given sentences in a target language mediated by latent alignments, sometimes constrained to be tree-structured (Wu, 1997). Models infer

these alignments jointly with a translation lexicon. However, they make strong conditional independence assumptions about the meanings of source tokens, and provide only one-to-many mappings between source and target tokens. While these word alignments may be used as a starting point for phrase-level extraction (Koehn et al., 2005; Chiang, 2007), they generally cannot be used when tokens are individually meaningless and non-alignable.

Most relevant to the current work, Faruqui and Dyer (2013) perform bilingual lexicon induction using parallel corpora by searching for words that share high mutual information. The approach we describe shares similar intuition but leverages general-purpose sequence models to enable context-sensitive alignment without requiring word-level correspondences.

Neural representations and predictions With the widespread use of neural network models for language processing, more recent approaches have derived alignments from predictions (or learned *representations*) rather than explicit generative models. For example, several approaches (Zenkel et al., 2020; Chen et al., 2021) use masked language models to learn word alignment by analyzing the contributions of source words in the prediction. Other works train multi-lingual models on parallel corpora, then extract alignments based on similarity of learned word representations in these models (Dou and Neubig, 2021).

Segmentation and translation In natural language, concepts are not always mappable to individual words. Often sub-word (morphemes) or super-word (phrases) segments encode basic units of meaning required for dictionary learning or trans-

lation. Performing alignment in these settings requires joint inference on both the segment boundary and its alignment. In this direction, Snyder and Barzilay (2008) describe a bilingual Bayesian model that learns to induce morpheme boundaries by marginalizing over all possible alignments. While the task was to learn morphological segmentation, a joint model of alignment and segmentation was used during training. In machine translation, Sennrich et al. (2015) study the problem of translating rare and unknown words by decomposing them into sub-word units using byte-pair encoding (BPE), a data compression algorithm that iteratively identifies frequent token sequences and replaces them with new tokens. Outside of multi-lingual settings, many probabilistic and information-theoretic approaches have been used to discover reusable sub-word units (Goldsmith, 2000; Smit et al., 2014; Bergmanis and Goldwater, 2017).

3 Approach

What do the various approaches to alignment described above have in common? In general, we expect a span \mathbf{x}_i to be aligned to a span \mathbf{y}_i if the two spans contain information about each other. In Fig. 1, it becomes easier to predict that one of the masked segments is *graph* knowing that the other masked segment is *graf*, and vice-versa. In fact, (*graph*, *graf*) is one of only a small number of pairs for which this is true: if we had instead masked (*graph*, *disco*), knowing the contents of one gap would not have made it any easier to predict the other one, because all requisite information would already be available in the context. Intuitively, *graph* and *graf* contain information about each other, while *graph* and *disco* do not.

This intuition can be formalized in terms of **pointwise mutual information (PMI)** (Fano, 1961). Given random variables \mathbf{X}_i and \mathbf{Y}_i , the PMI between two outcomes \mathbf{x}_i and \mathbf{y}_i is defined as:

$$\text{pmi}(\mathbf{x}_i; \mathbf{y}_i) = \log \frac{p(\mathbf{x}_i, \mathbf{y}_i)}{p(\mathbf{x}_i)p(\mathbf{y}_i)}. \quad (1)$$

$$= \log \frac{p(\mathbf{x}_i | \mathbf{y}_i)}{p(\mathbf{x}_i)} \quad (2)$$

Via Eq. (2), PMI may be understood as quantifying how much our confidence in the outcome \mathbf{x}_i increases after observing \mathbf{y}_i . This definition can also be extended to the conditional setting: given

some other random variable \mathbf{Z} , we may write:

$$\text{pmi}(\mathbf{x}_i; \mathbf{y}_i | \mathbf{z}) = \log \frac{p(\mathbf{x}_i, \mathbf{y}_i | \mathbf{z})}{p(\mathbf{x}_i | \mathbf{z})p(\mathbf{y}_i | \mathbf{z})} \quad (3)$$

In the context of alignment, if \mathbf{x}_i and \mathbf{y}_i are spans, and \mathbf{z} is the context in which they occur, \mathbf{x}_i and \mathbf{y}_i should be aligned precisely when their PMI (conditioned on \mathbf{z}) is large. INFOALIGN operationalizes this notion by first building a probabilistic model of source and target sequences, using this model to score spans based on conditional PMI, then uses scores to find the highest-scoring span alignments. Below, we describe each of these steps in more detail.

3.1 Masked Span Modeling

The first component of INFOALIGN is a joint probabilistic model of source and target spans in context. Let \mathbf{x}_i and \mathbf{y}_i be spans of sequences \mathbf{x} and \mathbf{y} . For convenience, let us define $\mathbf{x}_{-i} = \mathbf{x} \setminus \mathbf{x}_i$ (a version \mathbf{x} with the span \mathbf{x}_i masked out; see Fig. 1). We may define \mathbf{y}_{-i} correspondingly. Then, to compute the PMI between two spans in context (via Eq. (3)), we must compute the following three quantities:

$$p(\mathbf{x}_i, \mathbf{y}_i | \mathbf{x}_{-i}, \mathbf{y}_{-i}) \quad (4)$$

$$p(\mathbf{x}_i | \mathbf{x}_{-i}, \mathbf{y}_{-i}) \quad (5)$$

$$p(\mathbf{y}_i | \mathbf{x}_{-i}, \mathbf{y}_{-i}) \quad (6)$$

Each of these probability distributions is a kind of **masked language model** of a kind well-studied in the NLP literature: like the T5 and BART language models (Raffel et al., 2020; Lewis et al., 2019), all three quantities represent distributions over variable-length spans occurring in the middle of input sequences; like forgetful causal models (Liu et al., 2022), the latter two quantities mask multiple spans but predict only a subset. For large datasets, these distributions may be represented approximately using neural language models (Bengio et al., 2000). For small datasets, it is even possible to represent them using explicit frequency counts (Och and Ney, 2000). Indeed, it is possible to view Eqs. (4–6) as special kinds of *skip-gram* model (Huang et al., 1993) of a kind formerly popular in speech recognition and machine translation.

In practice, given a training set of paired sequences, we sample uniformly from the set of all maskings and train models to predict each of the three quantities above. We use encoder–decoder models, which generate \mathbf{x} , \mathbf{y} or both autoregressively (like T5 and BART) to avoid the indepen-

dence assumptions made by masked language models (like BERT).¹ As a concrete example, each term in the bottom right of Fig. 1 shows an example of an input–output pair used for training (or querying) these models. Inputs may contain [MASK], [HIDE] and [SEP] tokens, while outputs contain one prediction for each [MASK]ed span, delimited with [SEP] tokens if multiple [MASK]s are present.

3.2 Conditional PMI

Given models of Eqs. (4–6), we compute PMI exactly as in Eq. (3). As described below, it is useful to define one additional quantity, which we call the **cross-information** (CI):

$$\begin{aligned} \text{ci}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{x}_j, \mathbf{y}_j) = & \text{pmi}(\mathbf{x}_i; \mathbf{y}_i \mid \mathbf{x}_{-i}, \mathbf{y}_{-i}) \\ & + \text{pmi}(\mathbf{x}_j; \mathbf{y}_j \mid \mathbf{x}_{-j}, \mathbf{y}_{-j}) \\ & - \text{pmi}(\mathbf{x}_i; \mathbf{y}_j \mid \mathbf{x}_{-i}, \mathbf{y}_{-j}) \\ & - \text{pmi}(\mathbf{x}_j; \mathbf{y}_i \mid \mathbf{x}_{-j}, \mathbf{y}_{-i}) \quad (7) \end{aligned}$$

If \mathbf{x}_i and \mathbf{x}_j are adjacent spans (likewise \mathbf{y}_i and \mathbf{y}_j), then CI intuitively measures the quality of a *partition* of the aligned spans $[\mathbf{x}_i, \mathbf{x}_j]$ and $[\mathbf{y}_i, \mathbf{y}_j]$ into aligned sub-spans (Fig. 2). If CI is less than zero, then unaligned sub-spans contain as much or more information about each other compared to aligned spans. If there is no split of the two combined spans with positive CI, then those spans are not divisible further.

3.3 Unit Discovery

In some applications (like the reference resolution task we will study in Section 5), tools for computing PMI between arbitrary spans are useful even without producing a single span-level alignment between source and target sentences. But in other applications (like the word translation task in Section 4) explicit segment-to-segment alignments are useful, e.g. for building a lexicon of frequently aligned span pairs. Thus, the final component of INFOALIGN is an algorithm for constructing a hierarchical, span-level sequence-to-sequence alignment using the measures defined in Section 3.2.

This procedure is defined formally in Algorithm 1. It is broadly inspired by the splitting parser of Stern et al. (2017). We begin by assuming that

¹In the case of neural models, we cannot guarantee that Eqs. (5–6) will exactly correspond to marginals of Eq. (4), even though we expect them to do so asymptotically (Goyal et al., 2022; Hennigen and Kim, 2023). In experiments, even though neural models sometimes made “impossible” predictions (e.g. $p(\mathbf{x}_i, \mathbf{y}_i) > p(\mathbf{x}_i)$), we found this did not appear to limit their effectiveness at discovering high-quality alignments.

Algorithm 1 Alignment via top-down splitting

```

1: function ALIGN( $\mathbf{x}_i, \mathbf{y}_i$ )
2:   # Add current input to set of aligned spans.
3:   spans  $\leftarrow \{(\mathbf{x}_i, \mathbf{y}_i)\}$ 
4:   # Find the highest-scoring split.
5:    $a^*, b^* \leftarrow \arg \max_{a,b}$ 
6:      $\text{ci}(\mathbf{x}_i^{<a}, \mathbf{y}_i^{<b}, \mathbf{x}_i^{\geq a}, \mathbf{y}_i^{\geq b})$ 
7:   # If this split has non-positive C.I., stop.
8:   if  $\text{ci}(\mathbf{x}_i^{<a^*}, \mathbf{y}_i^{<b^*}, \mathbf{x}_i^{\geq a^*}, \mathbf{y}_i^{\geq b^*}) \leq 0$  then
9:     return spans
10:  # Otherwise, recurse on splits.
11:  spans  $\leftarrow$  spans  $\cup$  ALIGN( $\mathbf{x}_i^{<a^*}, \mathbf{y}_i^{<b^*}$ )
12:  spans  $\leftarrow$  spans  $\cup$  ALIGN( $\mathbf{x}_i^{\geq a^*}, \mathbf{y}_i^{\geq b^*}$ )
13:  return spans

```

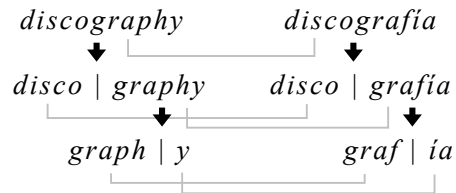


Figure 2: Alignment via top-down splitting. Beginning with complete (source, target) pairs, we recursively, synchronously split these pairs until their CI becomes non-positive.

the entire source sentence is aligned to the entire target sentence, then recursively *split* aligned spans into pairs of aligned sub-spans by maximizing CI. The procedure stops when no split yields positive CI. It runs in $\mathcal{O}(m^2n^2)$ time (where m and n are the lengths of \mathbf{x} and \mathbf{y} respectively). The version described in Algorithm 1 (and used in our experiments) assumes that alignments are monotonic, but can be easily extended to non-monotonic alignments (with only constant overhead) by also considering CI between pairs $(\mathbf{x}_i^{<a}, \mathbf{y}_i^{\geq b})$ and $(\mathbf{x}_i^{\geq a}, \mathbf{y}_i^{<b})$ on line 5.

Aside: exact alignment The above procedure may be viewed as greedily attempting to optimize an objective of the form:

$$\max_{\mathbf{A} \in \mathcal{A}} \sum_{(\mathbf{x}_i, \mathbf{y}_i, \mathbf{x}_j, \mathbf{y}_j) \in \mathbf{A}} \text{ci}(\mathbf{x}_i, \mathbf{y}_i, \mathbf{x}_j, \mathbf{y}_j) \quad (8)$$

where \mathcal{A} is the set of hierarchical alignments \mathbf{A} between \mathbf{x} and \mathbf{y} (e.g. the set depicted with gray lines in Fig. 2). While not used in our experiments, it is actually possible to optimize this quantity exactly using standard algorithms for forced alignment in

inversion–transduction grammars (Wu, 1997), with CI as a scoring function. This procedure requires $\mathcal{O}(m^3n^3)$ time (but only $\mathcal{O}(m^2n^2)$ evaluations of the scoring function).

In the remainder of this paper, we evaluate INFOALIGN on two different word learning problems: word-level MT and grounded color naming. Each is described below.

4 Experiments: Translation

Our first set of experiments focuses on learning to translate words (at the character level) by learning a morphological lexicon. In this task, models are trained set of inflected word pairs in source and target languages, and evaluated on their ability to translate novel word forms. Generalization of this kind is only possible with a correct model of the internal morphological structure of words:

4.1 Dataset

Our experiments focus on translating from English to Spanish. This language pair presents a particularly interesting case because Spanish is a *fusional* language: single, non-decomposable morphemes often carry information about number, person, tense and gender simultaneously. These may in turn interact with lemmas in complex ways. Spanish morphology is also in general more complex than English, so the learned mapping must be one-to-many. Thus, inferred morphological lexica must encapsulate information about morpheme pairs that may encode different pieces of information, and learned predictors must use morpheme-level information in a manner sensitive to global word structure.

We evaluate using word pairs from the MUSE project (Lample et al., 2017). In the training split, this dataset contains 11977 paired word forms, corresponding to 5000 unique English forms and 10166 unique Spanish forms. The test set, meanwhile, contains 2975 paired forms, with 1500 unique English inputs. However, at most 1046 of these are, even in principle, predictable on the basis of the training set (in the sense that they are expressible in terms of paired spans that co-occurred during training).

We evaluate performance on this task using two metrics. First, for the subset of words that are (in principle) exactly predictable, we report **exact match (E.M.)**: given an English input, does the model’s predicted Spanish output correspond

to *any* valid Spanish translation? Second, for all words (even those that cannot be translated exactly), we report **character edit distance (C.E.D.)**: the minimum Levenshtein distance between the predicted translation and any valid translation.

4.2 Model

To apply INFOALIGN to the word translation task, we first extract a dictionary of morpheme pairs from forced alignments, then compose these morphemes together using a neural sequence model.

Morpheme lexicon We use the procedure described in Section 3 to induce a joint segmentation and alignment of every word pair in the training set. We run Algorithm 1 up to a maximum depth of 2, in practice analyzing each word as a (prefix, suffix) pair or single morpheme. Surprisingly, we found that we obtained higher-quality predictions using exact count-based estimates of Eq. (3) rather than a neural model.²

We then construct a morpheme-level lexicon with one entry for each leaf (pair of aligned, non-decomposable segments) in the induced alignments. Each lexicon entry is assigned a score corresponding to the conditional PMI between the aligned segments. When a given pair of segments appears in multiple training words, we add these PMI-based scores together.

LM-guided decoding In parallel with morpheme extraction, we train an ordinary character-level sequence-to-sequence model (a single-layer, 1024-dimensional LSTM with attention, which we found more effective than any transformer variant we tried on the small training dataset; Hochreiter and Schmidhuber, 1997). Finally, given an input \mathbf{x} , we predict:

$$\max_{\substack{(\mathbf{x}_i, \mathbf{y}_i), (\mathbf{x}_j, \mathbf{y}_j) \\ \mathbf{x}_i \mathbf{x}_j = \mathbf{x}}} (\text{score}(\mathbf{x}_i, \mathbf{y}_i) + \text{score}(\mathbf{x}_j, \mathbf{y}_j) + \lambda \log p_{\text{LM}}(\mathbf{y}_i \mathbf{y}_j | \mathbf{x}_i \mathbf{x}_j)) \quad (9)$$

where morpheme pairs $(\mathbf{x}_i, \mathbf{y}_i)$ and $(\mathbf{x}_j, \mathbf{y}_j)$ are taken from the lexicon, and score denotes the entry score computed as described above.

4.3 Baselines

We compare INFOALIGN to several baselines:

²Because alignments are only computed on the training set, backoff methods are not needed to guarantee these models assign probability to all inputs on which they will be evaluated. Sparsity is a potential issue; while we use all counts exactly, future work might incorporate smoothing methods of the kind commonly used in n -gram models (Och and Ney, 2000).

	E.M. \uparrow	C.E.D. \downarrow
INFOALIGN	0.17	2.13
+ MORFESSOR	0.03	0.64
- context	0.15	2.62
- rescoring	0.14	2.39
SEQ2SEQ	0.15	4.52
+ MORFESSOR	0.07	2.48

Table 1: Evaluation results for the word translation task. E.M. denotes exact string match and C.E.D. denotes character edit distance; both are computed with respect to the best choice in the set of valid translations. The base INFOALIGN model outperforms a standard sequence-to-sequence model, with or without pre-tokenization using MORFESSOR. Both context-conditioning and rescoring with a sequence model are necessary to obtain these results.

- Ablations of the main INFOALIGN model: one of which removes **context** (computing PMI, rather than conditional PMI, between aligned spans), and one of which removes **rescoring** with the neural sequence model. These ablations evaluate the role of the specific decoding criterion described in Eq. (9).
- A **neural SEQ2SEQ** baseline that directly generates from the sequence model rather than using it for rescoring, with no lexicon-based scores or decoding constraints. This baseline evaluates the role of the learned lexicon in improving generalization performance.
- Variants of both INFOALIGN and SEQ2SEQ that operate not on characters, but on word pieces inferred using MORFESSOR (2.0), a classical (monolingual) morphological segmentation algorithm (Smit et al., 2014) that identifies frequently occurring spans using a minimum description length criterion. These variants evaluate the quality of segments discovered by INFOALIGN relative to other approaches to unsupervised segmentation.

4.4 Results

Table 1 shows results of our experimental evaluation. INFOALIGN outperforms SEQ2SEQ with and without MORFESSOR-based unit discovery; both rescoring and context are important for high-quality span alignment. Intriguingly, applying MORFESSOR to INFOALIGN substantially worsens exact match, but improves character edit distance.

Examples of discovered morphemes are shown in Table 2. They include frequently occurring stems

English	Spanish	Score
-s	-s	120.0
	-os	76.3
	-es	54.3
-ing	-ando	19.4
	-iendo	18.5
	-ndo	17.3
-ation	-ación	11.4
	-ción	8.6
	-aciones	2.9
-ed	-do	30.8
	-ó	19.2
	-da	11.7
publish-	edito-	2.0
	publica-	2.0
	editoria-	1.0
believ-	cre-	2.0

Table 2: Discovered word piece alignments in English–Spanish word translation. Only the 3 highest-scoring entries for each word are shown. Discovered correspondences include inflectional and derivational morphology, as well as lemmas. In some cases multiple translations are possible (e.g. English *-ed*, which can correspond to the past perfect, imperfect, or preterite in Spanish), and multiple lexicon entries are generated.

English	Spanish	INFOALIGN	SEQ2SEQ
<i>impression</i>	<i>impresión</i>	<i>impres-ión</i>	<i>presenta</i>
<i>relocated</i>	<i>trasladó</i>	<i>r-localizado</i>	<i>recariado</i>
<i>prisoner</i>	<i>prisionera</i>	<i>carcel-ador</i>	<i>respadar</i>
<i>grows</i>	<i>crece</i>	<i>crece-s</i>	<i>crece</i>
<i>keys</i>	<i>llaves</i>	<i>clave-s</i>	<i>claves</i>

Table 3: Example outputs from the INFOALIGN and SEQ2SEQ models. *Spanish* shows the (closest) ground-truth translation, while subsequent columns show model predictions. For INFOALIGN, morpheme boundaries are denoted with a -. INFOALIGN often generates correct translations; sometimes translations are phonotactically and semantically plausible even when incorrect.

and affixes, and reflect variability in allowed translation resulting from the many-to-many mapping between English and Spanish word forms. Table 3 shows model predictions that use these inferred alignments. Even when incorrect, these are often close (the English morpheme *re* is mapped to the Spanish span *r*, resulting in a phonotactically unacceptable prediction); in other cases, they are semantically plausible even when incorrect (*carcelador*, the model’s predicted translation of *prisoner*, is not a real word but could be reasonably translated as *jailer*). By contrast, the SEQ2SEQ model sometimes generates words with no obvious cor-

respondence to the input (*respadar*) or generates inflections that were seen in training data (*crece*).

5 Experiments: Reference Resolution

Our other experiments focus on a grounded reference resolution task. In this task, referring expressions are generated in a highly ambiguous perceptual context; at training time, learners must jointly infer word meanings and their context-dependent referents; at evaluation time, learners must resolve references for new inputs.

5.1 Dataset

We use the Colors in Context dataset from [Monroe et al. \(2017\)](#). Each example consists of a natural language referring expression paired with a set of three color patches (Table 4). To generate referring expressions, human annotators were shown the three patches and asked to refer to one of them; another annotator was then evaluated on their ability to correctly resolve the referent. Generated expressions are very sensitive to context (*redder of the two brownish colors, darker purple*).

Most work on Colors in Context has studied a supervised version of the problem, in which models learn to predict or resolve references given ground-truth information about the target color. In contrast, we evaluate on an *unsupervised* version of the reference resolution problem, in which learners do not have access to the target even at training time, and must jointly learn word meanings and their contextual referents. Colors were generated with constant luminosity but varying hue and saturation, so each color is presented to learners as a pair of integers.

As above, we use two metrics to evaluate predictors for this task. First, their exact match success at the reference game: what fraction of expressions was correctly resolved? Second, their perceptual distance: how far was the learner’s chosen color from the true color (measured in HSV space)?

5.2 Model

Rather than first extracting a fixed lexicon mapping names to color parts, we use computed PMI between utterances and single color patches to directly identify the referents of natural language expressions. We begin by training a model exactly as in Section 3 (learning to predict masked versions of all possible source/target spans). For these experiments, unlike above, we use a trained transformer to compute conditional PMI.

At evaluation time, we successively mask each candidate referent (a complete H, S pair), then compute its PMI with the (unmasked) input utterance conditional on the other candidate referents. Finally, we select the referent with the greatest PMI.

Why should we expect this procedure to work? Because referents in the colors in context dataset are context-sensitive, we expect targets to be predictable only given information about the other available referents. The scoring model thus needs to implement a version of pragmatic reference resolution internally (something that past work has found neural models capable of; [Monroe et al., 2017](#)) in order to assign high probability to contextually appropriate color descriptions.

5.3 Baselines

We compare INFOALIGN to:

- An ablation of the main INFOALIGN model, as in Section 4, that removes conditioning on context (and scores unconditional PMI between colors and referring expressions).
- A **neural attention** baseline. We concatenate (color, expression) pairs into single sequences, then train a masked language model on these sequences exactly as in the BERT model ([Devlin et al., 2019](#)). Finally, we predict by selecting the color having greatest *cross attention* with the input sequence, averaging over all heads and layers.

5.4 Results

Results are shown in Table 5. As above, INFOALIGN outperforms the standard neural baseline; here, even more than the translation task, conditional alignment is essential for good performance. The unsupervised version of this task is challenging, and performance remains far from perfect, but INFOALIGN performs significantly better than chance (in contrast to the attention model, which is only a few percentage points better than a chance baseline).

Examples of model predictions are shown in Table 4. INFOALIGN successfully resolves complex and context-dependent references, including examples containing comparatives (*redder, darker*), similes (*color of a cherry*) and even more complex uses of context (*combo of the other 2 colors*). In contrast, the attention-based scoring method often makes basic mistakes (choosing a bright green when the expression refers to *brownish colors*).



















Referring expression	A	B	C	G.T.	I.A.	M.A.
<i>it s a combo of the other 2 colors</i>				B	B	A
<i>color of a cherry</i>				B	B	B
<i>redder of the two brownish colors</i>				C	C	A
<i>the brightest pink</i>				C	C	A
<i>blue</i>				A	C	B
<i>well the darker purple</i>				B	A	B

Table 4: Example predictions on the Colors in Context task. Columns A, B and C show the candidate referents presented to the learner. G.T. shows the ground truth label (seen by the human annotator but not by models). I.A. shows predictions from INFOALIGN, while M.A. shows predictions from the MASKEDATTENTION model. INFOALIGN often makes correct predictions even when context is required to interpret expressions (as in the first line).

	E.M. \uparrow	C.D. \downarrow
INFOALIGN	0.50	49.0
– context	0.37	66.4
MASKEDATTENTION	0.34	77.4

Table 5: Evaluation results for the color reference resolution task. Only INFOALIGN performs significantly above chance, but succeeds only when context is used to compute alignment scores.

Performance, while above chance, remains significantly below the near-perfect accuracy that many supervised models achieve on this task; we expect that more sophisticated visual representations, or perhaps explicit pragmatic procedures of the kind described by Andreas and Klein (2016) or McDowell and Goodman (2019) might improve results.

6 Limitations

One major limitation of the proposed approach is runtime. Applying this method to extract a structured lexicon, as in Section 3.3, is computationally costly, especially in the presence of deeper structures than investigated here. Extracting these correspondences requires more effort than inspecting the behavior of a (quadratic-time) attention mechanism.

Additionally, PMI can only be computed if we have the ability to assign a *normalized* probability to a masked sequence. Outside of language domains, many of today’s most sophisticated generative models (including GANs and diffusion models) define intractable probability distributions, meaning that additional modeling work will be required to scale INFOALIGN to these more complex domains (e.g. images).

7 Conclusion

We have presented INFOALIGN, an information-theoretic approach to alignment that can identify context-dependent, span level correspondences between inputs in multiple modalities. INFOALIGN outperforms both classical unit discovery and neural sequence modeling approaches in both word translation and reference resolution domains. More broadly, INFOALIGN offers a new approach for thinking about what an alignment *is* in domains where the primitive elements of alignment (analogous to words in machine translation) are unknown, and complete source \rightarrow target generative models cannot be specified. By deriving alignments from information-theoretic measures, we can use the modern neural sequence modeling toolkit to obtain meaningful correspondences between data of diverse types.

References

- Jacob Andreas and Dan Klein. 2016. Reasoning about pragmatics with neural listeners and speakers. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*.
- Yoshua Bengio, Réjean Ducharme, and Pascal Vincent. 2000. A neural probabilistic language model. In *Advances in Neural Information Processing Systems*.
- Toms Bergmanis and Sharon Goldwater. 2017. From segmentation to analyses: a probabilistic model for unsupervised morphology induction. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.
- Peter F Brown, John Cocke, Stephen A Della Pietra, Vincent J Della Pietra, Frederick Jelinek, John Lafferty, Robert L Mercer, and Paul S Roossin. 1990. A statistical approach to machine translation. *Computational linguistics*, 16(2):79–85.

- Chi Chen, Maosong Sun, and Yang Liu. 2021. Mask-Align: Self-supervised neural word alignment. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- David Chiang. 2007. [Hierarchical phrase-based translation](#). *Computational Linguistics*, 33(2):201–228.
- 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 Conference of the North American Chapter of the Association for Computational Linguistics*.
- Zi-Yi Dou and Graham Neubig. 2021. [Word alignment by fine-tuning embeddings on parallel corpora](#). In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.
- Robert M Fano. 1961. Transmission of information: A statistical theory of communications. *American Journal of Physics*, 29(11):793–794.
- Manaal Faruqi and Chris Dyer. 2013. An information-theoretic approach to bilingual word clustering. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- John Goldsmith. 2000. Linguistica: An automatic morphological analyzer. In *Proceedings of 36th Meeting of the Chicago Linguistic Society*.
- Kartik Goyal, Chris Dyer, and Taylor Berg-Kirkpatrick. 2022. Exposing the implicit energy networks behind masked language models via metropolis–hastings. In *Proceedings of the International Conference on Learning Representations*.
- Lucas Torroba Hennigen and Yoon Kim. 2023. Deriving language models from masked language models. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Leo Hickey. 1998. *The pragmatics of translation*. Multilingual Matters.
- Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.
- Xuedong Huang, Fileno Allewa, Hsiao-Wuen Hon, Mei-Yuh Hwang, Kai-Fu Lee, and Ronald Rosenfeld. 1993. The SPHINX-II speech recognition system: an overview. *Computer Speech & Language*, 7(2):137–148.
- Philipp Koehn, Amittai Axelrod, Alexandra Birch Mayne, Chris Callison-Burch, Miles Osborne, and David Talbot. 2005. [Edinburgh system description for the 2005 IWSLT speech translation evaluation](#). In *Proceedings of the International Workshop on Spoken Language Translation*.
- Guillaume Lample, Alexis Conneau, Ludovic Denoyer, and Marc’Aurelio Ranzato. 2017. Unsupervised machine translation using monolingual corpora only. In *Proceedings of the International Conference on Learning Representations*.
- Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Hao Liu, Xinyang Geng, Lisa Lee, Igor Mordatch, Sergey Levine, Sharan Narang, and Pieter Abbeel. 2022. FCM: Forgetful causal masking makes causal language models better zero-shot learners. *arXiv preprint arXiv:2210.13432*.
- Bill McDowell and Noah Goodman. 2019. [Learning from omission](#). In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Will Monroe, Robert XD Hawkins, Noah D Goodman, and Christopher Potts. 2017. Colors in context: A pragmatic neural model for grounded language understanding. *Transactions of the Association for Computational Linguistics*, 5:325–338.
- Richard Montague. 1973. The proper treatment of quantification in ordinary English. In *Approaches to natural language: Proceedings of the 1970 Stanford workshop on grammar and semantics*.
- Franz Josef Och and Hermann Ney. 2000. [A comparison of alignment models for statistical machine translation](#). In *Proceedings of the International Conference on Computational Linguistics*.
- Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2020. Exploring the limits of transfer learning with a unified text-to-text transformer. *Journal of Machine Learning Research*, 21(1):5485–5551.
- Rico Sennrich, Barry Haddow, and Alexandra Birch. 2015. Neural machine translation of rare words with subword units. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Peter Smit, Sami Virpioja, Stig-Arne Grönroos, Mikko Kurimo, et al. 2014. Morfessor 2.0: Toolkit for statistical morphological segmentation. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*.
- Benjamin Snyder and Regina Barzilay. 2008. Unsupervised multilingual learning for morphological segmentation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.
- Mitchell Stern, Jacob Andreas, and Dan Klein. 2017. A minimal span-based neural constituency parser. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.

- Bill Thompson, Sean Roberts, and Gary Lupyan. 2018. Quantifying semantic similarity across languages. In *Proceedings of the Annual Conference of the Cognitive Science Society*.
- Dekai Wu. 1997. Stochastic inversion transduction grammars and bilingual parsing of parallel corpora. *Computational linguistics*, 23(3):377–403.
- Thomas Zenkel, Joern Wuebker, and John DeNero. 2020. End-to-end neural word alignment outperforms GIZA++. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*.