A Bayesian Conditional Model for Sense Induction and Translation

Abstract

We introduce a Bayesian conditional model of translation with latent variables representing alignments, topics, and source language word senses. The model structure and priors are motivated by insights from sense discrimination, topic modeling, and translation modeling. The conditional structure of the model means the topics and senses explain regularities in how words translate, rather than in the distribution of word types. Using the inferred word alignments and translation probabilities, we show gains on German–English alignment and translation tasks, and we also obtain state-of-the-art results on an unsupervised English word sense induction task by incorporating Czech translations of the training data.

1 Introduction

Lexical translation models (Brown et al., 1993; Vogel et al., 1996; Dyer et al., 2013) continue to play a central role in statistical machine translation. They are used for word alignment and to smooth distributions for more complex translation models (Koehn et al., 2003; Chiang, 2007), as reranking features (Och et al., 2004), and for bilingual lexicon induction. A central simplifying assumption in these models is that the translation decision at each position in the output is conditionally independent of all other decisions, given the identity of the aligned source word. Such models therefore fail to explicitly capture the translation variants of polysemous and homonymous words. This paper propose a new translation model with latent word sense indicators that addresses this limitation.

We motivate the high level structure of our model with an example. Consider the French word *prix*. This word has two primary senses corresponding to English *price* and *prize*. In a French–English parallel corpus, we will find this homonymy reflected by a relatively "flat" empirical distribution over the English translations used for this word. However, if we considered the sentences discussing a topic like finance, we expect to find that the first sense is meant far more often, and that the translation distribution is far more peaked around words meaning *price*. On the other hand, in the sentences about sports, we expect to find the second sense more often, with the translation distribution peaked around *prize*.

Our proposed model (§2) captures this interaction between topic, sense, and translation. We use priors encode four insights that are well-known from the literature on word sense discrimination, topic modeling, and translation: (i) that different senses of the same word will tend to translate differently (Diab and Resnik, 2002) (Yao et al., 2012); (ii) that words will tend to have a small number of frequent translations (Riley and Gildea, 2012); (iii) that topics (i.e., contexts) will have few senses per word type—the "one sense per discourse" heuristic (Gale et al., 1992b); and (iv) that documents will tend to have a small number of topics (Blei et al., 2003). Although each of these insights is appreciated, our work is the first to combine them all into one single model where they interact. Additionally, the conditional structure of the model means notions of topic and word sense will be used to explain how words translate, rather than the distribution of words in a monolingual corpus. Given a parallel corpus, posterior inference (§3) lets us reason about a variety of useful marginal quantities: translation probabilities in

the context of a given sense, the number of word senses for a word type, sense labels, and word alignments.

We evaluate our model on two separate tasks (§4). The first is a German–English bitext alignment task, which we evaluate both intrinsically, using Alignment Error Rate (AER), and extrinsically, by introducing a feature based on our model into the translation system. The second is an English word sense induction task, for which the training data was independently translated into Czech.

2 Model

Since our model incorporates elements of a probabilistic topic model (Blei et al., 2003) and a lexical translation model (Brown et al., 1993) using hierarchical Pitman–Yor process priors (Pitman and Yor, 1997; Teh, 2006) (Goldwater et al., 2011), we briefly review these (§2.1) and then describe our complete model (§2.2).

2.1 Background

The Pitman–Yor Process. The Pitman–Yor process (PYP) is a generalization of the nonparametric Dirichlet Process (DP). In addition to the strength α and base distribution P_0 of a DP, the PYP has a third parameter $d \in [0, 1)$ called the discount parameter. The DP is a special case of the PYP process where $d = 0$, and there is a roughly exponential decay in the probability of new draws from the base distribution. Larger values of d dampen the "rich get richer" dynamic of the DP resulting in the "heavier" tail of a power-law distribution. Such distributions are effective models of type distributions in natural language (Teh, 2006) (Goldwater et al., 2006).

Posterior inference is possible by marginalizing the draws from PYPs using a Chinese Restaurant process (CRP) sampling scheme. In contrast to the DP version of the CRP, the number of customers seated at a table is "discounted" by d. Thus, tables with just one or two customers contribute much less than in a DP as $d \to 1$.

Latent Dirichlet Allocation. Latent Dirichlet Allocation (LDA) is a probabilistic topic model that models a collection of documents using shared topics which are distributions over words. Documents are generated by choosing a mixture of topics θ_d, and then the ith word is generated by selecting a topic $t_{d,i}$ from θ_d and then word $w_{d,i}$ from ϕ_t. θ_d is drawn from a Dirichlet distribution, which encodes the belief that documents will generally be about a small number of topics. The ϕ_k are either set to maximize likelihood or inferred as draws from a Dirichlet distribution. Thus this model is able to infer a distribution over topics for each document in a collection, as well as a list of words that most strongly indicate membership to each topic.

One-Parameter Model 2. IBM Model 2 (Brown et al., 1993) is a generative model of translation that is usually used to infer word alignments in a bitext. Given a source sentence $\boldsymbol{f} = \langle f_1, f_2, \ldots, f_m \rangle$ and a target length n, the model generates a sequence of alignments $\boldsymbol{a} = \langle a_1, a_2, \ldots, a_n \rangle$ and target sentence $\boldsymbol{e} = \langle e_1, e_2, \ldots, e_n \rangle$. The alignment distribution we use is due to Dyer et al. (2013), and says that $p(a_i = j \mid m, n) \propto \exp -\lambda \left| \frac{i}{n} - \frac{j}{m} \right|$. In this model, alignment decisions for each target position i are independent of each other, making extremely efficient inference possible. Finally each e_i is generated conditional on f_{a_i}.

2.2 The Topic-Sense-Translation Model

Our model is based on several key intuitions. First and foremost is that different senses of polysemous words will translate differently into a foreign language. For example, the English word *plant* is polysemous—it can mean either a green leafy lifeform or a place where goods are manufactured. German uses completely different words these two senses: *Pflanze* for the green thing and *Anlage* for the one with smokestacks. The second intuition behind our model is that within a given domain, polysemous words tend to have a dominant sense. For example, within the domain of ecology, the biological sense of plant will be used almost exclusively. Conversely, when speaking about commerce, the manufacturing sense of plant will be dominant. Our third intuition is that each document in a corpus tends to pertain to be about a small number of topics. That is to say that when reading a news article about worker conditions in manufacturing plants, we can feel relatively sure that the author will not switch gears and talk about ecology later in the article. Of course, this can happen, we just expect *a priori* that documents will be about few topics. Since polysemous words tend to exhibit one sense per topic, and each document typically pertains to just one topic, the

T		(Observed)	Number of topics
Z		(Observed)	Number of senses per word
ρ		(Observed)	Parameter for senses' Geometric prior
V_E		(Observed)	Number of target vocabulary types
θ_0	$\sim$	$\mathrm{PYP}(\mathrm{Uniform}(T))$	Base distribution over topics
ψ_f	$\sim$	$\mathrm{PYP}(\mathrm{TruncGeom}(\rho, Z))$	Distribution over senses for source type f
$\psi_{f,t} \mid \psi_f$	$\sim$	$\mathrm{PYP}(\psi_f)$	Distribution over senses for source type f in topic t
ϕ_0	$\sim$	$\mathrm{PYP}(\mathrm{Uniform}(V_E))$	Distribution of target words
$\phi_f \mid \phi_0$	$\sim$	$\mathrm{PYP}(\phi_0)$	Translation table for source type f
$\phi_{f,z} \mid \phi_f$	$\sim$	$\mathrm{PYP}(\phi_{f,z})$	Translation table for source type f with sense z
θ_d	$\sim$	$\mathrm{PYP}(\theta_0)$	Distribution over topics for document d
n_d		(Observed)	Number of source tokens in document d
m_d		(Observed)	Number of target tokens in d
$f_{d,j}$		(Observed)	Source word j
$t_{d,j}$	$\sim$	$\mathrm{Categorical}(\theta_d)$	Topic of source word j
$z_{d,j} \mid t_{d,j}, f_{d,j}$	$\sim$	$\mathrm{Categorical}(\psi_{f_{d,j},t_{d,j}})$	Sense of source word j
$a_{d,i} \mid n_d, m_d$	$\sim$	$\mathrm{Diagonal}(i, n_d, m_d)$	Alignment link for target word i
$e_{d,i} \mid \boldsymbol{a}, \boldsymbol{f}$	$\sim$	$\mathrm{Categorical}(\phi_{f_{a_{d,i}},z_{a_{d,i}}})$	Target word i

Figure 1: Formal definition of our model. The first group of variables are given, the second group are parameters marginalized during inference, and the third are observed or inferred data values. PYP hyperparameters are not shown here, but discussed in the text.

combination of intuition number two and intuition number three gives rise to the famous "one sense per discourse" maxim. We capture these intuitions as priors. Importantly, we only these to hold in expectation, they may be violated in individual cases.

The formal generative process shown in Figure 1 and the corresponding plate diagram in Figure 2. We describe the process here in text and explain our modeling decisions. We assume that the parallel corpus contains T topics, and that no word type has more than Z, and that the number of senses for a word type is governed by a truncated geometric distribution with parameter ρ. For each parallel document d in the corpus, we draw mixing proportions θ_d over the T topics from a PYP with a uniform base distribution of T topics. A document about the effect of oil drilling on river ecosystems might have been generated from a θ_d that assigns 80% of its probability mass to the topic corresponding to ECOLOGY, with the remaining 20% assigned to the topic corresponding to ECONOMY. Then, for each *source word* f_i in d, we draw its topic indicator $t_{i,d}$ from θ_d. Note that in contrast to more familiar LDA models, we do not generate f_i, but rather observe it, since our intention is to let topic variables explain translation regularities, not to explain the distribution of source word types.

For each (f, t) tuple of a source word type and a topic, we generate draw a distribution $\psi_{f,t}$ over sense indicators (i.e., natural numbers in $[1, Z]$) as a draw from a hierarchical PYP with the truncated geometric base distribution described above. We have already predicted a topic for each source word instance in our corpus, so we can use this distribution $\psi_{f,t}$ to assign each instance a sense indicator z. The distribution $\psi_{f,t}$ might tell us, for example, that if we see the word *bank* paired with the ECOLOGY topic, it will refer to a sloping land mass at the side of a river 95% of the time. If, however, we see *bank* with the topic FINANCE, it refers to the financial institution sense 75% of the time, and the act of hoarding 20% of the time.[1]

Next, we draw a per-sense translation table $\phi_{f,z}$ for each (f, z) tuple of a source word and a sense. To illustrate the behavior of $\phi_{f,z}$ we would expect that in a German–English model when $f =$ *bank* and z is the majority sense of the ECOLOGY topic,

[1]The decision to use separate topic and sense variables may seem somewhat surprising. However, our intuition is that topics are distributions over *word senses*, and that different word senses will be shared across topics.

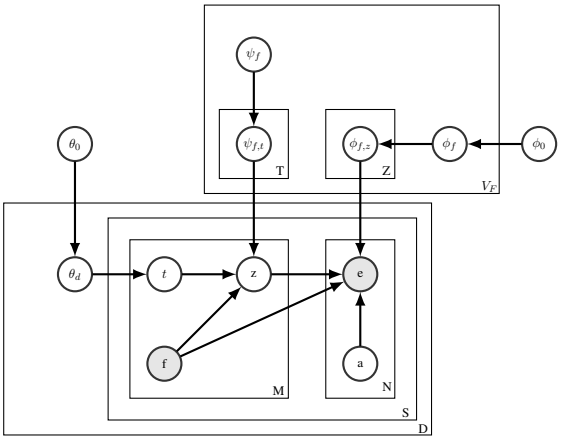

Figure 2: Plate diagram of our model. Explanation of notation and distribution information is given in Figure 1.

$\phi_{f,z}$ would likely assign a high probability to the German word *Ufer*.

At this point, we have generated sense tags for every source word in the corpus, and have sense-specific translation parameters. We now generate alignments using the one-parameter Model 2 process described above and then generate each target word conditioning on the sense-tagged source word and using the appropriate $\phi_{f,z}$.

Remarks on the PYP Hierarchy. To make our generative model more robust to noisy and sparse data, we use hierarchical priors over topics for each document, θ_d, the distribution of senses given a word type and a topic $\psi_{f,t}$, and the per-sense translation table $\phi_{f,z}$. This hierarchical structure encourages these key distributions to be "close" to an underlying distribution that conditions on one less variable. Thus, in cases of noise or sparsity, related distributions can share statistical strength, giving improved estimates of the probability of rare phenomena.

In particular, we desire that the per-sense translation tables be based upon an underlying translation table that conditions only on the source word, ignoring the sense. This structure allows each per-sense translation table to then specialize, assigning more probability mass to a select few translations at the expense of others. For example, the underlying sense-agnostic translation table of the English word *plant* might say that it translates to the German *Anlage* 60% of the time and *Pflanze* 30% of the time, with the last 10% reserved for other words. The translation table of $plant_1$ might give the translation *Anlage* extra probability, leaving the percentages at *Anlage* 90%, *Pflanze* 5%, with the final 5% forming the tail. The translation table of $plant_2$ will hopefully choose to reassign the probability mass quite differently, choosing something like *Pflanze* 80% of the time. Since PYP processes encourage reuse of existing translations, this will be a stable state if these two senses really do account for how *plant* translates in the data. In this way, the translation tables are free to specialize and choose only a few translations out of the palette provided by the underlying distribution, yet still must be somewhat similar, thus preventing rampant overfitting and diminishing the probability of using of completely implausible English words.

We use such a structure for θ_d, with a prior θ_0 that does not condition on any particular document, thus representing the frequency with which a topic will be used in some document. For $\psi_{f,t}$, we do something similar for, backing off to a ψ_f that conditions on f, but not on t, thus representing the prior probability that f will be used with a particular sense in a new topic. For the sense-specific translation probability, we first back off to a distribution ϕ_f that conditions only on the source word f, but not on its sense z. That back off, in turn, backs off to yet another prior ϕ_0, which conditions on neither f nor z, thus representing only the prior probability of an English word being used in a sense-agnostic translation model.

Hyperparameters. The proposed model has a large number of hyperparameters, i.e., the discount and concentration parameters for each PYP. Rather than specifying these directly, we use a so-called "vague" priors on the hyperparameter values marginalize them as well. For discount parameters, we draw from $\mathrm{Beta}(1,1)$ (i.e., the uniform distribution on $(0,1)$), and for concentration parameters, we draw from $\mathrm{Gamma}(1,1)$, which encourages values closer to 0 (Johnson and Goldwater, 2009).

To ensure that ψ_f is sparse, representing the intuition that most words have a very small number of senses, we impose a Geometric prior on ψ_f. This ensures that the model will prefer to limit each word to a single sense, unless there is substantial evidence of polysemy. Furthermore, the computational cost of each additional sense per word increases exponentially, ensuring that the model will not use 5 senses when two will do.

3 Inference with Gibbs Sampling

Given a corpus of parallel data, we reason about posterior distributions over senses and alignments using a collapsed Gibbs sampler. In particular, the θ, ψ, and ϕ variables are all collapsed out (Chen et al., 2011). All experiments are run with 1000 iterations of the resulting Gibbs sampler. All relevant outputs are calculated by taking the average of outputs after each iteration, with a burn-in of 500 iterations. Pitman–Yor priors have their parameters re-sampled using slice sampling (Neal, 2003) every 30 iterations. All latent categorical variables are initialized to values drawn from a uniform distribution over the relevant domain.

For our experiments, we cap the possible number of senses per word at 5, which seems reasonable as only extremely few words have more than 5 general senses. Furthermore, we limit the number of topics to 5. This is in contrast to many unsupervised topic models, such as LDA (Blei et al., 2003), which typically set the number of topics to several hundred. We can use such a small number of topics only because our topics do not directly generate words, as in LDA, but are instead only concerned with differences in how words translate. This means that our topics do not necessarily correlate well with the human notion of topic, but still capture the correlation between senses of different words. Future work will look at ways to marginalize these variables.

If, for example, two human senses, say sports and manufacturing, do not oppositely affect the sense probabilities of any individual word, putting them into the same cluster will have no effect on results. For example, knowing that we're in the manufacturing domain means that the word "produce" is likely a verb meaning to generate output, and not fresh vegetables. If we're talking about sports, however, it is unlikely that we will see the word "produce" at all. So long as the set of words that that have manufacturing specific senses is disjoint from the set of words that have sports specific senses, combining the two into one pseudo-topic has no negative impact on the efficacy of the model.

4 Experiments

We tested the effectiveness of our model on several tasks, including both intrinsic and extrinsic evaluations of the alignments produced by our model, the effect of introducing the probabilities generated by our model as a feature in a translation system, and the quality of the monolingual word senses induced by our model.

	Prec.	Rec.	AER
fast_align (`dir`)	71.1%	73.8%	27.6%
fast_align (`gd`)	72.4%	75.8%	26.0%
fast_align (`gdfa`)	70.8%	76.6%	26.6%
this work (`dir`)	71.7%	78.6%	25.2%
this work (`gd`)	**77.0**%	80.7%	**21.3**%
this work (`gdfa`)	75.3%	**81.9**%	21.8%

Table 1: Intrinsic Alignment Quality Results

4.1 Intrinisic Alignment Evaluation

First, we tested the intrinsic effectiveness of the model as a bilingual aligner. We compare the alignments produced on a data set to human annotated alignment links using the standard metrics, Precision, Recall, and Alignment Error Rate (AER), and compare to Dyer et al. (2013)'s fast_align, a strong, modern baseline. We compare both directional word alignments (dir), and bidirectional alignments symmetrized with the grow-diag (gd) and grow-diag-final-and (gdfa) heuristics (Och and Ney, 2003) on a $100,000$-segment subset of the German–English news commentary data from WMT. As a preprocessing step, we run Dyer et al. (2010)'s tokenizer on both the German and English sides of the data. We then run their German compound splitter on the German side. Finally, we run each side of the data through the Snowball Stemmer (Porter, 2002) for its language. Examples of latent senses in this data uncovered by our model can be seen in Figure 3.

In the directional case we see only moderate gains in precision, but nearly 5% better recall using our model, leading to about 2.5% lower AER. After symmetrizing, the results are even more dramatic: a 4.5% increase in precision, and just over 5% improvement in recall, leading to over 4.5% improvement in AER using both symmetrization heuristics, as seen in Table 1.

4.2 Extrinsic Alignment Evaluation

Furthermore, we can test the effectiveness of these improved alignments in an actual translation system. To this end, we built a hierarchical phrase-based translation (Chiang, 2007) German–English translation system using $153,800$ segments ($3,947,916$ German tokens) of WMT news commentary data. We build the system using the

Kopf$_1$	*Kopf*$_2$	*Kopf*$_3$
10.9%	38.2%	50.9%
capita	down	head
minds	upside	heads
it	capita	mind
heads	leaders	minds
	head	cool

schlagen$_1$	*schlagen*$_2$
21.1%	78.9%
beat	are
defeat	proposing
suggest	capitalize
propose	hit
capitalize	propose

Figure 3: Examples of senses uncovered by our model for the German words *Kopf* ("head") and *schlagen* ("beat/hit" or "propose"), their inferred relative frequencies, and their most probable translations.

cdec decoder and framework (Dyer et al., 2010). The data were aligned by using either fast_align or our model and symmetrized with grow-diag-final-and. The system was tuned with 20 iterations of Batch MIRA (Cherry and Foster, 2012) on the WMT 2012 test set, and tested against the WMT 2013 test set. Overall, our alignments yielded an improvement of +0.6 BLEU over the baseline alignments produced by fast_align. Full results can be seen in Table 2.

	BLEU	METEOR	TER	$\lvert\hat{e}\rvert/\lvert e^*\rvert$
fast_align	18.7	37.7	**59.1**	93.6
this work	**19.3**	**38.4**	**59.1**	**95.7**

Table 2: Extrinsic alignment quality results (German–English translation).

4.3 Translation Reranking

The above tests test the effectiveness of the alignments output by the model, but have yet to show the full power of the model as a feature in translation. To that end, we included our aligner's scores as a feature in a set of k-best re-ranking experiments.

We use the same baseline German–English translation system as the previous experiment and extract 1000-best lists on the tuning and testing sets. We then compute the log likelihood of each hypothesis given each source segment under the posterior predictive distribution of our sense-augmented alignment model. In these reranking experiments, the alignments produced by fast_align were used.

If we were to include this feature inside a decoder, rather than by adding it to a k-best list to be reranked, we would lack the full target side of the sentence, making computation of the log likelihood under our model difficult. Furthermore, even in our case where we do have access to target language hypotheses, the latent variables in our model cannot be topologically sorted for easy inference. For both of these reasons, it is attractive to build a monolingual sense induction model.

To this end, we build a logistic regression classifier trained on the data provided by our German–English training set. For each token t in a sentence S in the source side of the training corpus, we construct an input feature vector wherein each feature is a tuple of source word types, (w, v). The feature (w, v) takes the value 1 iff $w = t$ and $v \in S$, i.e. if w is the type of the token t and v co-occurs with t in the sentence S, and takes the value 0 otherwise. In place of gold standard data, we train our model on the entirety of the training set and use the inferred senses as "wood standard" labels. This classifier is able to recover the senses of its training data with 95.3% accuracy.

With this classifier, we can easily label the sense of each token on the source side of our k-best list. With the senses observed, the optimal alignment under our model is easily computable, along with its score. It is the log of this score that is added as a feature to the k-best list.

	BLEU	METEOR	TER	$\lvert\hat{e}\rvert/\lvert e^*\rvert$
Baseline	18.7	37.7	**59.1**	93.6
Our Model	**19.0**	**38.0**	59.5	**96.0**

Table 3: Results of re-ranking experiments using three standard MT evaluation metrics. The last column shows the length of the translations relative to the reference length.

We then use a discriminative reranker based on the loss function described by Yadollahpour et al. (2013), trained on our held-out tuning set to produce an updated set of weights, including a weight for the feature representing our model's score. We find that adding the log of the likelihood our model assigns to each element in the k-best list as a fea-

ture results in an improvement of +0.3 BLEU over our baseline system. Detailed results can be found in Table 3.

4.4 Sense Induction

Finally, we examine how well the senses inferred by our model correspond to human word sense intuitions by testing it on an English sense induction task. For this task, we used the the section of the *OntoNotes* corpus (Hovy et al., 2006) taken from the Wall Street Journal, consisting of $49,208$ segments ($1,214,801$ English tokens), and often used for sense induction tasks (Brody and Lapata, 2009) (Yao and Van Durme, 2011). Our approach requires, however, the rather non-conventional constraint that the training data has to have been translated into a foriegn language. Thankfully, Čmejrek et al. (2004) translated this section of the Wall Street Journal into Czech, as part of their larger project, the Prague Czech–English Dependency Treebank.

While running the full model is certainly tractable, we can achieve substantial speed up, and reduce overfitting, by capitalizing on an extra benefit of our use of a Gibbs sampler to do inference. Gibbs sampling affords us the ability to use any amount of supervised data available by simply setting the value of the relevant latent variables to the supervised values, while continuing to sample the other, unsupervised latent variables. Since in this task we are interested in the word senses rather than the alignments, we decided to align the corpus with fast_align, and use the resulting alignments as supervised data. We then run the model, and output the model's predicted word sense for each English token.

Due to the fact that the senses output by our model are in no way ordered, the sense numbers output by our model are not guaranteed to match up with the sense numbers used by the OntoNotes annotators. As such, when evaluating our model's output, we first build a map for the model's hypothesized word senses to the OntoNotes reference word senses, by choosing the reference sense r that maximizes $p(r \mid w, h)$ for each word w and hypothesis sense h. We then transform our algorithm's output deterministically, using this map, into a form that is directly comparable to the OntoNotes annotations.

Previous work, such as (Brody and Lapata, 2009), has built this map by holding out a development set to learn the map from the hypothesis sense space to the gold standard sense space. Since we have limited data in Czech, and are loath to hold out extra data, we instead decided to use the leave-one-out methodology to map our output to the reference sense space. For each word in the corpus, we compute the empirical relative frequencies $f(r \mid w, h)$ from all the other word instances in the corpus, and using that model, we predict the label the reference will assign to the one given word.

Using this method, we find that our algorithm predicts the sense of English words with 85.5% accuracy, compared to the 80.9% accuracy achieved by the naïve baseline of always choosing the most frequent sense for each English word, 86.7% achieved by Yao and Van Durme (2011), and 87.3% achieved by Brody and Lapata (2009). For comparison, we also trained our monolingual classifier to predict English word senses monolingually, using its output on the Wall Street Journal task and achieving 84.5% accuracy. The resulting monolingual logistic regression classifier performed reasonably, but slightly worse than the full bilingual model.

Our bilingual model's performance, as seen in Table 4, did not quite match the current state of the art evaluated in terms of matching a monolingual sense inventory. Since word sense inventories are typically constructed to reflect monolingual semantic differences rather than to distinguish translation decisions (which will be influenced by numerous factors, but typically correlate well with different senses), we look at the performance of our model in which sense-specific distributions generate monolingual context words rather than translations. Using Brody & Lapata's highest performing feature category, a 10-word context window around each source word, when resampling a the sense z for a source word f_m, we consider the probability of generating the nearby words, $p(f_{m+i} \mid f_m, z)$ for $i \in [-10, 10], i \neq 0$, likewise assuming these distributions are draws from a nonparametric prior. Adding in this one category of monolingual features increased the performance of our model by 3.2%, allowing us to surpass the current reported state-of-the-art reported in Brody and Lapata (2009).

Method	Accuracy
One Sense Per Type	80.9%
Monolingual	84.5 %
This work (Bilingual)	85.5%
Yao and Van Durme (2011)	86.7%
Brody and Lapata (2009)	87.3%
This work (Mono context)	**88.7%**

Table 4: Sense induction performance.

5 Related Work

Topic-based domain adaptation of the translation model was previously explored by Tam et al. (2007), who propose a bilingual topic model based on Latent Semantic Analysis. Similarly Eidelman et al. (2012) use Latent Dirichlet Allocation to form topics over source sentences. Ruiz and Federico (2011) apply topic models based on Probabilistic Latent Semantic Analysis to the adaptation of language models. The insight that adaptation of translation distributions is better captured with a "conditional" topic model is also explored in the work of Hasler et al. (2014), who model this topic-specific models of phrasal translation probabilities (rather than the word translation probabilities we use).

More broadly, a polylingual implementation of topic models is given by Mimno et al. (2009). They detail a generative model capable of finding cross-lingual topics given parallel documents in two or more languages, and apply their model to many down stream tasks, including translation.

Lau et al. (2012) monolingually use topic models to infer senses, paying particular attention to detection of novel word senses and tokens of these new "emergent" meanings.

Zhao and Xing (2006) describe a topic-aware alignment model quite similar to that used in this work. However, they limit sentences to have only a single topic, rather than being represented by a mixture of topics as in this work. Furthermore, they apply their model directly as a word aligner, achieving improved word alignment accuracy but failing to demonstrate improved translation quality.

The idea of re-ranking hypotheses using forward (target given source) and reverse (source given target) IBM Model 1 scores was first explored in Och et al. (2004). They found that reverse Model 1 scores were their top performing feature, while forward scores showed no noticable gain.

Gale et al. (1992a) were among the first to leverage large corpora to do automatic word sense disambiguation. Brody and Lapata (2009) took a modern, Bayesian approach to the problem to nice effect. Yao and Van Durme (2011) continued this path by using non-parametric Bayesian models to tackle the problem, but still limit themselves to a monolingual feature set.

Diab and Resnik (2002) applied clustering techniques to translation tables to infer latent senses from parallel data.

Lyse (2011) explore a large variety of techniques for exploiting differences in how lexical items translate to infer word senses, specifically for the English–Norwegian language pair.

Task-specific representation learning for language, which is one way of understanding this work, is becoming increasingly common as neural networks grow in popularity (Socher et al., 2013) (Kalchbrenner and Blunsom, 2013) (Devlin et al., 2014); however, these have been largely in the form of distributed representations, rather than the discrete topic and sense indicators we infer in this work.

6 Conclusion

In this paper we proposed a novel technique for handling vocabulary items that exhibit domain-sensitive polysemy by using a Bayesian model to simultaneously infer topics, word senses, and translation tables from a bitext. We evaluated the resulting alignments both intrinsically, using AER, and extrinsically, as part of a full translation system. Finally, we show that our model can be used to achieve good performance in an unsupervised word sense induction task, provided that the training data has been translated into another language, and that state-of-the-art performance can be achieved by combining our model with the power of source-side co-occurrence information.

Acknowledgments

References

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *JMLR*.

Samuel Brody and Mirella Lapata. 2009. Bayesian word sense induction. In *Proc. of EACL*, pages 103–111.

Peter F. Brown, Vincent J. Della Pietra, Stephen A. Della Pietra, and Robert L. Mercer. 1993. The

mathematics of statistical machine translation: Parameter estimation. *Computational Linguistics*.

Changyou Chen, Lan Du, and Wray Buntine. 2011. Sampling table configurations for the hierarchical poisson-dirichlet process. In *Machine Learning and Knowledge Discovery in Databases*.

Colin Cherry and George Foster. 2012. Batch tuning strategies for statistical machine translation. In *Proc. of NAACL*.

David Chiang. 2007. Hierarchical phrase-based translation. *Computational Linguistics*.

Jacom Devlin, Rabih Zbib, Zhongqiang Huang, Thomas Lamar, Richard Schwartz, and John Makhoul. 2014. Fast and robust neural network joint models for statistical machine translation.

Mona Diab and Philip Resnik. 2002. An unsupervised method for word sense tagging using parallel corpora. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*.

Chris Dyer, Adam Lopez, Juri Ganitkevitch, Johnathan Weese, Ferhan Ture, Phil Blunsom, Hendra Setiawan, Vladimir Eidelman, and Philip Resnik. 2010. cdec: A decoder, alignment, and learning framework for finite-state and context-free translation models. In *Proc. of ACL*.

Chris Dyer, Victor Chahuneau, and Noah A Smith. 2013. A simple, fast, and effective reparameterization of IBM model 2. In *Proc. of NAACL*.

Vladimir Eidelman, Jordan Boyd-Graber, and Philip Resnik. 2012. Topic models for dynamic translation model adaptation. In *Proc. of ACL*.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992a. A method for disambiguating word senses in a large corpus. *Computers and the Humanities*.

William A. Gale, Kenneth W. Church, and David Yarowsky. 1992b. One sense per discourse. In *Proceedings of the workshop on Speech and Natural Language*.

Sharon Goldwater, Tom Griffiths, and Mark Johnson. 2006. Interpolating between types and tokens by estimating power-law generators.

Sharon Goldwater, Thomas L Griffiths, and Mark Johnson. 2011. Producing power-law distributions and damping word frequencies with two-stage language models. *The Journal of Machine Learning Research*.

Eva Hasler, Phil Blunsom, Philipp Koehn, and Barry Haddow. 2014. Dynamic topic adaptation for phrase-based MT. In *Proc. of EACL*.

Eduard H. Hovy, Mitchell P. Marcus, Martha Palmer, Lance A. Ramshaw, and Ralph M. Weischedel. 2006. Ontonotes: The 90% solution. In *Proc. of NAACL*.

Mark Johnson and Sharon Goldwater. 2009. Improving nonparameteric bayesian inference: experiments on unsupervised word segmentation with adaptor grammars. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent convolutional neural networks for discourse compositionality. *Proceedings of the 2013 Workshop on Continuous Vector Space Models and their Compositionality*.

Philipp Koehn, Franz J. Och, and Daniel Marcu. 2003. Statistical phrase-based translation. In *Proc. of NAACL*.

Jey Han Lau, Paul Cook, Diana McCarthy, David Newman, and Timothy Baldwin. 2012. Word sense induction for novel sense detection.

Gunn Inger Lyse. 2011. *Translation-based Word Sense Disambiguation*. Ph.D. thesis, The University of Bergen.

David Mimno, Hanna M. Wallach, Jason Naradowsky, David A. Smith, and Andrew McCallum. 2009. Polylingual topic models.

Radford M Neal. 2003. Slice sampling.

Franz Josef Och and Hermann Ney. 2003. A systematic comparison of various statistical alignment models. *Computational Linguistics*.

Franz J. Och, Daniel Gildea, Sanjeev Khudanpur, Anoop Sarkar, Kenji Yamada, Alex Fraser, Shankar Kumar, Libin Shen, David Smith, Katherine Eng, Viren Jain, Zhen Jin, and Dragomir Radev. 2004. A smorgasbord of features for statistical machine translation. In *Proc. of NAACL*.

Jim Pitman and Marc Yor. 1997. The two-parameter Poisson-Dirichlet distribution derived from a stable subordinator. *The Annals of Probability*.

Martin F. Porter. 2002. Snowball: A language for stemming algorithms.

Darcey Riley and Daniel Gildea. 2012. Improving the IBM alignment models using Variational Bayes. In *Association for Computational Linguistics (ACL) short paper*.

Nick Ruiz and Marcello Federico. 2011. Topic adaptation for lecture translation through bilingual latent semantic models. In *Proc. of WMT*.

Richard Socher, Alex Perelygin, Jean Y Wu, Jason Chuang, Christopher D Manning, Andrew Y Ng, and Christopher Potts. 2013. Recursive deep models for semantic compositionality over a sentiment treebank. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Yik C. Tam, Ian Lane, and Tanja Schultz. 2007. Bilingual-LSA based LM adaptation for spoken language translation. In *Proc. of ACL*, pages 520–527.

Yee Whye Teh. 2006. A hierarchical Bayesian language model based on Pitman–Yor processes. In *Proc. of ACL*.

Martin Čmejrek, Jan Hajič, and Vladislav Kuboň. 2004. Prague Czech–English dependency treebank: Syntactically annotated resources for machine translation. In *Proc. of EAMT*.

Stephan Vogel, Hermann Ney, and Christoph Tillmann. 1996. HMM-based word alignment in statistical translation. In *Proc. of COLING*.

Payman Yadollahpour, Dhruv Batra, and Gregory Shakhnarovich. 2013. Discriminative re-ranking of diverse segmentations.

Xuchen Yao and Benjamin Van Durme. 2011. Nonparametric Bayesian word sense induction.

Xuchen Yao, Benjamin Van Durme, and Chris Callison-Burch. 2012. Expectations of word sense in parallel corpora. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*.

Bing Zhao and Eric P. Xing. 2006. BiTAM: bilingual topic admixture models for word alignment. In *Proc. of ACL*.