Improving Ontology Translation from Disentangled Semantic and Language Representations

Mingjie Tian[1], **Fausto Giunchiglia**[2], **Hao Xu**[3,*]
[1] School of Artificial Intelligence, Jilin University
[2] Department of Information Engineering and Computer Science, University of Trento
[3] College of Computer Science and Technology, Jilin University
mjtian19@mails.jlu.edu.cn, fausto.giunchiglia@unitn.it, xuhao@jlu.edu.cn

Abstract

Ontology is the basis of knowledge representation, and it is necessary to translate ontologies that are normally expressed in English into other languages in order to achieve exchange across languages. Building a domain-specific translation system is essential due to the extremely focused words used and the inadequacy of contextual information. In this paper, we introduce disentangled representations under cross-lingual agreement to alleviate the aforementioned issues. We introduce semantic and language representations and integrate extra losses to induce disentangled representations that capture different information. To reduce the gap between the ontology label and the hypothesis generated by the translation model, we further integrate adversarial learning. In order to guide the generation of translation candidates, the semantic matching strategy is incorporated into the decoding phase. Experiments on the four English-to-German ontologies of different domains show that the proposed method achieves improvements over the baselines.

Keywords: ontology translation; disentangled representation; cross-lingual agreement, adversarial learning; fusion decoding

Introduction

Through a shared understanding of a conceptualization of a domain, ontology underpin the underlying representation of the knowledge (Gruber, 1995). Ontologies provide an institutional framework for the digital archive, with the concepts ruled as bibliometrics of key terms that reflect domain-specific notions. The utilization of ontologies has proved advantageous for numerous applications, such as in the areas of information extraction (Buitelaar, Cimiano, Frank, Hartung, & Racioppa, 2008), semantic search (Fernandez et al., 2008), and natural language generation (Bontcheva, 2005). Despite this extensive usage, the majority of ontologies have only been described in English. These monolingual resources must be converted into multilingual equivalents in order to make ontological information available beyond language borders and to benefit users of other languages.

Ontology labels often have a high degree of domain specificity, and parallel resources can not provide linguistic background for ontologies. Only 15% of the lexical items can be detected in the general parallel resources, according to our preliminary ontology localization attempt, while the other 85% are absent. Ontology labels are typically only short text pieces that differ linguistically from free text, and it is challenging for labels composed of just a few words to provide adequate context to encourage neural machine translation (NMT) systems to translate the labels to a specific domain. Building a domain-specific translation system with reliable translation candidates is necessary for ontology translation due to the extremely specialized vocabulary and absence of contextual information.

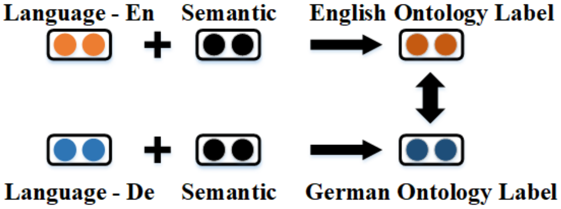


Figure 1: A motivation example for disentangling the ontology label to semantic and language variables.

Some previous studies integrate the transfer learning paradigm into the machine translation model to improve ontology translation. The basic model is initially trained on a substantial general-domain parallel corpus, and it is then fine-tuned on aligned parts from related domain datasets or ontologies. The majority of the approaches shift the domain to a specific one by filtering sentences from generic corpora or by incorporating domain-relevant external resources (Arcan, Turchi, & Buitelaar, 2015; McCrae et al., 2016; Arcan, Torregrosa, & Buitelaar, 2017). Although the inference of subwords may be somewhat alleviated in most circumstances when particular terminologies in ontology labels rarely exist in parallel corpora or fine-tuned data, the inference of translation candidates frequently cannot be learned effectively. At the same time, when translating the short format text through pre-trained models trained on long ones, translation errors frequently happen as a consequence of the absence of contextual information support.

We introduce disentangled representations under cross-lingual agreement, as shown in Fig. 1, to overcome the aforementioned issues that occur when NMT translates ontology labels. Firstly, the equivalent signals exhibited in the parallel data enrich the bidirectional contextual information for the words on both language sides and accelerate the word alignment procedure in NMT (Luong, Pham, & Manning, 2015a; Hermann & Blunsom, 2014). Additionally, word embeddings are employed by NMT systems as latent features for

[*] Corresponding Author

In M. Goldwater, F. K. Anggoro, B. K. Hayes, & D. C. Ong (Eds.), *Proceedings of the 45th Annual Conference of the Cognitive Science Society*.

representing parallel texts, even though it does not explicitly model language discrimination information. To disentangle various attributes of ontology labels, as shown in Fig. 1, two distinct types of representation could be considered: the semantic meaning and the language discrimination. The semantic representation is designed to encapsulate the meaning of ontology labels independent of languages. In contrast, the language representations should represent categories of the presented language (e.g., English, German). Partitive units in disentangled representations are those that are responsive to shifts in a particular component (Bengio, Courville, & Vincent, 2013), and it has been demonstrated that adopting such representations is favorable for generalization and interpretability (Achille & Soatto, 2018).

We propose a method for disentangling semantic and linguistic factors in order to improve their interpretability and discriminative effects during translation processes. We introduce two continuous latent variables to capture semantic meaning and language discrimination, and we learn these representations by optimizing the evidence lower bound (ELBO) with a Variational Autoencoder (VAE)-like (Kingma & Welling, 2014; Rezende, Mohamed, & Wierstra, 2014) approach. In order to improve the disentanglement of the learned representations, we integrate extra losses in our method that are intended to induce the latent representations to capture different information. At the same time, to reduce the difference between the ontology label and the translation generated by the NMT model, we integrate the semantic representations with adversarial learning during the training of the NMT. Additionally, the semantic matching strategy is incorporated into the log-likelihood during the decoding phase to guide the generation of translation candidates. We evaluate the proposed method for ontology translation from English to German in four domains. Experimental results show that the method achieves improvements over the baselines, demonstrating the effectiveness of exploiting disentangled representations under cross-lingual agreement for NMT. Further analysis also shows that the disentangled representations transfer learned knowledge to the NMT model.

Proposed Method

We describe our method for improving ontology translation through disentangled representations. The overall framework of our proposed method is shown in Fig. 2.

Disentangled Representations

Our method extends the vanilla VAE (Kingma & Welling, 2014) by adopting two distinct latent variables, z_{sem} and z_{lang}, to capture semantic and language information, respectively (shown in Fig. 3). We assume that the probability of an ontology label x could be computed as follows:

$$\begin{aligned} p(x) &= \int p(z_{sem}, z_{lang})\, p(x|z_{sem}, z_{lang})\, dz_{sem} dz_{syn} \\ &= \int p(z_{sem})\, p(z_{lang})\, p(x|z_{sem}, z_{lang})\, dz_{sem} dz_{syn} \end{aligned} \quad (1)$$

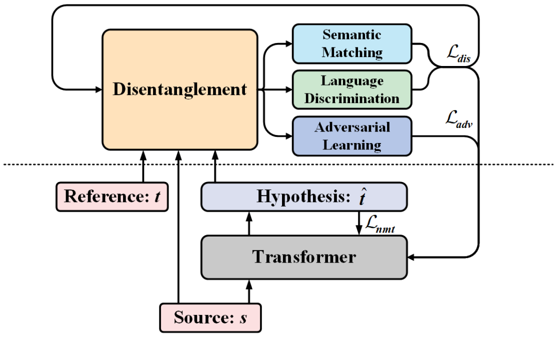


Figure 2: Diagram of proposed method.

where $p(z_{sem})$ and $p(z_{lang})$ are the priors; both are set to be independent multivariate Gaussian $N(0, I)$.

We optimize the ELBO for training:

$$\begin{aligned} \log p(x) \geq & \mathop{\mathbb{E}}_{q(z_{sem}|x)q(z_{lang}|x)} [\log p(x|z_{sem}, z_{lang})] \\ & - \beta_{KL}^{sem} KL(q(z_{sem}|x)||p(z_{sem})) \\ & - \beta_{KL}^{lang} KL(q(z_{lang}|x)||p(z_{lang})) = \text{ELBO} \end{aligned} \quad (2)$$

where $q(z_{sem}|x)$ and $q(z_{lang}|x)$ are posteriors for the semantic and language latent variables, respectively. We assume these two posteriors are independent, and taking the distribution of $N(\mu_{sem}, \sigma^2_{sem})$ and $N(\mu_{lang}, \sigma^2_{lang})$.

In the inference phase, motivated by the research line of phrase embeddings (Yazdani, Farahmand, & Henderson, 2015), we obtain the initial representation of the ontology label in two ways: 1) by averaging word representations (S. Wang & Zong, 2017), and 2) by using a recursive autoencoder (J. Zhang, Liu, Li, Zhou, & Zong, 2014), which is representative in terms of whether it is aware of the word orders while composing the phrase from component words. The representation of the ontology label x is fed into a feedforward neural network, which produces mean and variance of $q(z_{sem}|x)$ and $q(z_{lang}|x)$.

In the generation phrase, z_{sem} and z_{lang} are sampled through the reparameterization trick (Kingma & Welling, 2014), Then, these latent variables are concatenated as $z = [z_{sem}; z_{lang}]$ and fed into the generative model $p(x|z_{sem}, z_{lang})$ to generate the reconstruction of x. Finally, the reconstruction of x is input to the feedforward neural network to predict the bag of words of ontology labels.

Multi-Task Learning

In order to improve the quality of disentangled representations, we integrate the extra training objectives described below.

Semantic Matching. It has been demonstrated in past research (Luong, Pham, & Manning, 2015b; Kuang, Li, Branco, Luo, & Xiong, 2018) that cross-lingual alignment is beneficial for machine translation. Our goal is to propose a measurement for cross-lingual agreement that is based on the semantic meaning equivalence between translations. Firstly,

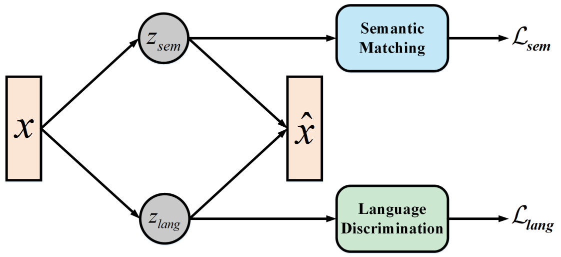


Figure 3: Diagram showing the disentanglement with semantic matching and language discrimination objectives.

we define the semantic distance of ontology label translation pairs (x, y) on different sides as:

$$\mathcal{L}_{dist}(x,y) = \frac{1}{2}\|z^x_{sem} - z^y_{sem}\|^2 \tag{3}$$

where z^x_{sem} and z^y_{sem} denote the disentangled semantic representations of the ontology labels x and y.

Ideally, we want the ontology label semantic distance for the positive examples to be much closer than that for the negative example. We enhance the semantic error with positive and negative examples, and the corresponding max-margin loss becomes:

$$\begin{aligned}\mathcal{L}_{margin}(x,y,y^*) = max\{0, &\mathcal{L}_{dist}(x,y)\\ &-\mathcal{L}_{dist}(x,y^*)+1\}\end{aligned} \tag{4}$$

where y^* denotes the negative sample sampled from the target language side and (x, y^*) denotes the negative pairs of ontology labels. We simply choose the most similar ontology labels in a mini-batch (other than those in the given translation pair) to generate negative samples for semantic matching.

Finally, the overall semantic matching objective becomes:

$$\mathcal{L}_{sem} = \mathcal{L}_{margin}(x,y,y^*) + \mathcal{L}_{margin}(y,x,x^*). \tag{5}$$

Language Discrimination. In the parallel dataset, each ontology label is presented in its own languages. We design a language-oriented training objective that correctly discriminate the language category through the disentangled language representations. Our language discriminator is parameterized by a two-layer feedforward neural networks $f(\cdot)$ with input from the disentangled language representation z_{lang}. Specifically, the language discrimination objective is defined as follows:

$$\mathcal{L}_{lang} = \mathop{\mathbb{E}}_{q(z_{lang}|x)}[-\sum_i \log \text{softmax}(f(z^i_{lang}))]. \tag{6}$$

Adversarial Learning. We introduce an adversarial learning mechanism between the NMT and the disentanglement module to enhance the translation effect. We view the NMT as a generator in the adversarial learning framework, while the disentanglement module is viewed as a discriminator. The purpose of training the NMT in adversarial learning is to generate translation candidate $\hat{y}$ that are closer to the source language ontology label x than ground truth y. We view the hypothesis, $\hat{y}$, as a positive example and the reference, y, as a negative example, giving the hypothesis generated by the NMT greater credibility than ground truth. The constructed adversarial learning objective is

$$\mathcal{L}_{adv} = max\{0, \mathcal{L}_{dist}(x,\hat{y}) - \mathcal{L}_{dist}(x,y) + 1\}. \tag{7}$$

Intuitively, updating the NMT parameters to minimize $\mathcal{L}_{adv}$ can be seen as learning to generate a translation, "cheating" the disentanglement module into believing that this translation should have a higher score than the corresponding ground-truth.

Training

The disentangled representations integrated into the machine translations are divided into two phases. Firstly, we pre-train the disentanglement module through

$$\begin{aligned}\mathcal{L}_{dis} =& \mathcal{L}_{vae} + \mathcal{L}_{aux}\\ =& -\text{EBLO} + \beta_{\text{sem}}\mathcal{L}_{\text{sem}} + \beta_{\text{lang}}\mathcal{L}_{\text{lang}} + \frac{\lambda}{2}\|\theta\|^2\end{aligned} \tag{8}$$

where hyper-parameters β_{sem}, β_{lang}, and β^{sem}_{KL}, β^{lang}_{KL} in Eq (2) control the strengths of each loss for the objective. Also, the regularization term $\frac{\lambda}{2}\|\theta\|^2$ is introduced to reduce overfitting during the phase of pre-training the disentanglement module.

Then, the NMT module and the disentanglement module are fine-tuned together with the following objective:

$$\mathcal{L} = \mathcal{L}_{nmt} + \beta_{dis}\mathcal{L}_{dis} + \beta_{adv}\mathcal{L}_{adv} + \frac{\lambda}{2}\|\theta\|^2 \tag{9}$$

where the hyper-parameters β_{dis} and β_{adv} control the strengths of disentanglement loss and adversarial loss for the final objective. The parameters θ in our model can be divided into three sets: 1) parameters of disentanglement of representations θ_{vae}; 2) parameters of language discrimination θ_{lang}; and 3) parameters of the NMT model θ_{nmt}.

Decoding

Integrating a language model in the decoding phase of machine translation has been proven to increase adaptation performance in previous studies (Stahlberg, Cross, & Stoyanov, 2018; Saunders, Stahlberg, & Byrne, 2019). In the decoding phase of machine translation, we combine disentanglement module and the NMT. When predicting the i-th position in the decoding phase, the probability can be calculated as

$$\begin{aligned}P(y_i|y_{<i},\boldsymbol{x}) = softmax(&S_{nmt}(y_i|y_{<i},\boldsymbol{x})\\ &+\beta_D S_{dis}(x,y_{<i}))\end{aligned} \tag{10}$$

where $S_{nmt}(y_i|y_{<i},\boldsymbol{x})$ denotes the output of the NMT projection layer without softmax. The hyper-parameter β_D balances the NMT output distribution and alignment score output by the disentanglement module. $S_{dis}(x,y_{<i})$ denotes the alignment score between x and $y_{<i}$, and the score is calculated as:

$$S_{dis}(x,y_{<i}) = log({z^x_{sem}}^\top z^{y_{<i}}_{sem}). \tag{11}$$

where $z^{y_{<i}}_{sem}$ denotes the disentangled representation of words preceding the i-th position.

Table 1: Generic and ontology datasets statistics.

Dataset	Lines	En Words	En Vocab.	En Avg. Length	De Words	De Vocab.	De Avg. Length
Generic	3,724,585	123,219,992	307,886	33.08	117,748,487	735,675	31.61
ICD	1,839	9,661	1,780	5.25	9,236	2,145	5.02
IFRS	2,757	24,871	998	9.02	26,197	1,621	9.5
STW	7,151	14,898	3,928	2.08	9,835	7,045	1.37
TheSoz	10,731	22,080	5,736	2.05	26,776	6,010	2.49

Experiment

Datasets and Settings

Considering that the majority of ontologies are represented in English, under the existing conditions, we study the translation from English to German, and dataset statistics are shown in Table 1.

Generic Datasets. We merged a number of parallel corpora to create the general domain parallel corpus, including JRC-Acquis (Steinberger et al., 2006), Europarl (Koehn, 2005), DGT (Steinberger et al., 2014), MultiUN Corpus (Eisele & Chen, 2010) and TED Talks (Cettolo, Girardi, & Federico, 2012).

Ontology Datasets. We verified the effectiveness of our method in four ontologies from different domains: 1) The International Classification of Diseases (**ICD**) ontology (The World Health Organization, 2011); 2) The International financial reporting standards (**IFRS**) ontology (Van Greuning, Scott, & Terblanche, 2011); 3) The **STW** ontology for economics (Borst & Neubert, 2009); 4) The Thesaurus for the Social Sciences (**TheSoz**) (Zapilko, Schaible, Mayr, & Mathiak, 2013).

We compare our method with the following baselines: 1) **RNN**-based NMT (Bahdanau, Cho, & Bengio, 2015) which extending the original encoder-decoder and adding an attention mechanism; 2) **Transformer** (Vaswani et al., 2017) which propose NMT based solely on attention mechanisms, dispensing with recurrence and convolutions entirely; 3) **VNMT** (B. Zhang, Xiong, Su, Duan, & Zhang, 2016a) which propose a variational model to learn this conditional distribution for neural machine translation; and 4) **AgreementNMT** (Yang et al., 2019) which propose a sentence-level agreement to minimize the difference between translation sentences.

We adopt byte pair encoding (Sennrich, Haddow, & Birch, 2016) with 32K merges to segment words into subword units. Our experiments demonstrate five-fold cross-validation. We evaluate the proposed approaches on our reimplemented Transformer model, which following the setups of the base model. Word embeddings are 512-dimensional and initialized randomly. For the disentanglement module, we apply an Adam (Kingma & Ba, 2015) optimizer with $b_1 = 0.9$ and $b_2 = 0.99$ and a base learning rate of 10^{-4}. The mini-batch size is 256, and the dropout rate is set to 0.1. The initial representations of the ontology label are 256-dimensional, and the dimension of each latent variables(z_{sem} and z_{lang}) is 128-dimensional. We set negative example size $k = 50$ for semantic matching in disentanglement module.

Table 2: BLEU scores of English-German ontology translation.

Model	ICD	IFRS	STW	TheSoz
RNN	14.99	17.99	14.21	13.98
Transformer	15.85	21.36	16.88	16.13
VNMT	15.81	17.39	15.43	14.63
AgreementNMT	16.78	21.49	16.82	16.34
WordEmbedd+AVG	**17.41**	21.57	**17.22**	**16.99**
WordEmbedd+RAE	16.24	21.08	16.90	16.34
TransEmbedd+AVG	16.80	**22.18**	16.81	16.49
TransEmbedd+RAE	17.06	21.76	16.87	16.21

We adopt the top-layer output of the Transformer's encoder and decoder (**TransEmbedd**) and the word embeddings (**WordEmbedd**) as the original representation. NMT learns the initial pre-training model on the generic datasets, and the disentanglement module is pre-trained in the ontology development set. In the phase of fine-tuning, the NMT module and disentanglement module continue training on the ontology evaluation set. Finally, both modules participated in the decoding on the test set. During the decoding phase,we apply with the beam size of 10 for beam search. The translation results are measured in case-insensitive BLEU (Papineni, Roukos, Ward, & Zhu, 2002). We select the hyperparameter value with the lowest $\mathcal{L}_{dis}$ and $\mathcal{L}$ on the validation set, which the validation set sampled from all ontology datasets. These values were tuned by grid search, but due to the large hyperparameter space, we gradually varied the hyper-parameters from 0.05 to 0.5 with an increment of 0.05 in each step. Our model achieved the best performance when $\beta_{KL}^{sem} = 1.0$, $\beta_{KL}^{lang} = 1.0$, $\beta_{sem} = 0.3$, $\beta_{lang} = 0.35$ and $\lambda = 0.05$ for pretraining disentanglement module. We obtain best performance for finetuning disentanglement module and NMT when $\beta_{dis} = 0.5$, $\beta_{adv} = 0.15$ and $\lambda = 0.05$. The best translation result is obtained when $\beta_D = 0.25$ during the decoding phase.

Translation Result

We report the experimental results on ontology translation in this section. With the choice of representations of words, we could adopt word embeddings(**WordEmbedd**) and the output

Table 3: Ablation study by adding variational autoencoder, semantic matching, language discrimination, adversarial learning and fusion decoding. $\mathcal{L}_{dis}$ indicates the integrating of $\mathcal{L}_{vae}$, $\mathcal{L}_{sem}$ and $\mathcal{L}_{lang}$.

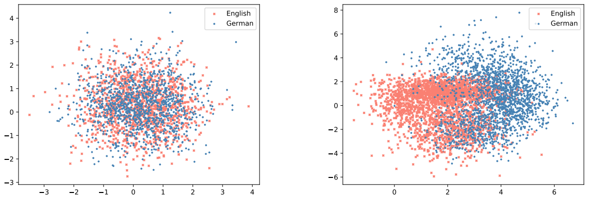
	ICD	IFRS	STW	TheSoz
$\mathcal{L}_{nmt}$	15.85	21.36	16.88	16.13
$\mathcal{L}_{nmt}, \mathcal{L}_{vae}$	14.57	18.02	13.17	14.09
$\mathcal{L}_{nmt}, \mathcal{L}_{vae}, \mathcal{L}_{sem}$	15.89	20.76	16.90	16.54
$\mathcal{L}_{nmt}, \mathcal{L}_{vae}, \mathcal{L}_{lang}$	14.87	20.21	16.01	16.21
$\mathcal{L}_{nmt}, \mathcal{L}_{dis}$	16.92	21.55	17.18	16.74
$\mathcal{L}_{nmt}, \mathcal{L}_{dis}, \mathcal{L}_{adv}$	17.04	21.48	16.99	16.76
ALL+Decoding	17.41	21.57	17.22	16.99

of the Transformer's encoder and decoder (**TransEmbedd**). Also, there are two candidates for the choice of learning initial representations of ontology labels: 1) averaging word representations (**AVG**); and 2) recursive auto-encoder (**RAE**). For the best performance, we explored every possible option, and Table 2 shows the performances measured in terms of the BLEU score on the baseline and our methods.

Among the baseline methods, Transformer has the best results on ICD and STW datasets, and Agreement NMT has the best results on IFRS and TheSoz datasets. The **WordEmbedd+AVG** option in our proposed method yields the best results across three datasets, whereas the **TransEmbedd+AVG** performs best on a single dataset. The **WordEmbedd+AVG** option's result exceeds baseline methods on all datasets. In the choice of composition method for initial representations for ontology labels, **AVG** has a better translation performance than **RAE**, due to the prediction of the generation model in the disentanglement module disregarding the word orders. Finally, these results imply that NMT benefits from disentangled representations during the translating and decoding phases.

Ablation Study

We conducted an ablation study and show the results in Table 3. We train our model from **Transformer** to **WordEmbedd+AVG** by gradually increasing training objectives in order to investigate the component's contributions. We can find that the performance cannot be improved by adding the VAE module alone (line 2); on the contrary, it will have the opposite effect. And the effect increases when the semantic matching objective (line 3) is added to the VAE, because the cross-lingual agreement constraint forces the VAE module to learn the equivalent relation. A significant raise in translation performance is observed by integrating the full disentanglement module(line 5), which essentially disentangles the latent variables into semantic and language spaces and transfers the learned knowledge to the NMT in a multi-task learning context. Additionally, the disentangled representations also improve translation performance during the decoding phase (line 7), which guides the generation of translation candidates. Finally, by incorporating all these modules, we achieve the best translation results.



(a) Semantic Space

(b) Language Space

Figure 4: t-SNE plots of the disentangled semantic and language spaces on ICD dataset.

Table 4: Alignment scores of ontology translation pairs.

	ICD	IFRS	STW	TheSoz
Transformer	1.23	1.45	0.98	1.01
WordEmbedd+AVG	40.34	32.92	45.31	46.31
WordEmbedd+RAE	24.34	30.12	41.97	42.97
TransEmbedd+AVG	27.54	29.29	40.39	33.39
TransEmbedd+RAE	34.01	28.43	41.54	40.54

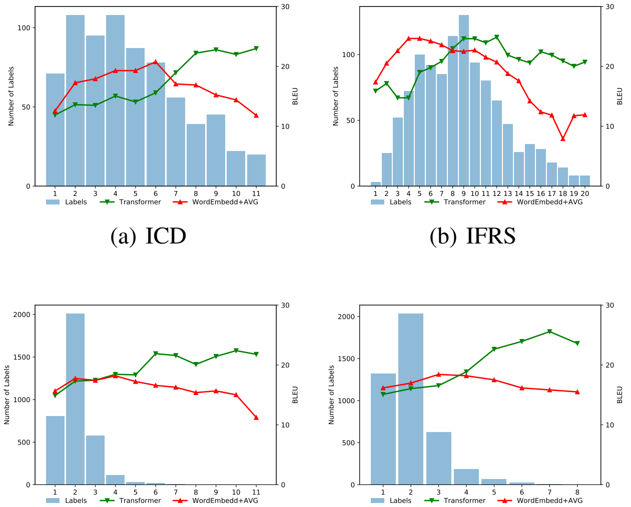
Disentangled Representations

We examine the disentangling quality of learned semantic and language of representations, primarily studying the latent space of ICD ontology datasets. We select ontology label translation pairs and visualize their latent representation in Fig. 4 via t-SNE plots (van der Maaten & Hinton, 2008). The red and blue points respectively represent the English and German ontology labels. We adopt z_{sem} and languagez_{lang} as semantic and language representations, and these vectors are induced by the **WordEmbedd+AVG** option in our method. The left side of the figure is the semantic representation space, in which the distribution of latent variables is coincident across languages. The right side of the figure shows the language representation space, which is well separated into two parts with different colors. It supports the claim that our method learns a discriminative language space. These results indicate the disentangling effect of our method, as the language space contains language discrimination information, whereas the semantic space maintains equivalence across languages.

Additionally, as shown in Table 4, we calculate the ontology label alignment score between translation ontology labels in order to further investigate the behavior of the disentangled semantic representations. The cosine similarity of the

Table 5: Changes in alignment score during pre-training and fine-tuning.

	ICD	IFRS	STW	TheSoz
pre-train dis	59.22	40.43	63.36	53.33
Transformer	1.23	1.45	0.98	1.01
w/ dis(Frozen)	25.29	15.82	30.21	27.68
w/ dis(Fine-tuning)	40.34	32.92	45.31	46.31



(a) ICD

(b) IFRS

(c) STW

(d) TheSoz

Figure 5: Ontology labels length distribution, and translation performances (BLEU) on each length.

semantic representation z_{sem} is used to determine the alignment scores. Our method produces a significant increase in alignment scores compared to the Transformers due to the disentangled semantic representations and semantic-specific training objectives. Since the Transformer does not explicitly model the alignment relation, its scores are incredibly low.

Further, as shown in Table 5, we investigate the alignment score changes that took place during pre-training and fine-tuning. We adopt the alignment scores based on the **WordEmbedd+AVG** option, and the result show that the strongest alignment ability is possessed by the pre-trained disentanglement module. It's interesting to note that when the NMT is fine-tuned, the alignment ability of the disentangled representation declines while the settings of the disentanglement modules are frozen(w/ dis(Frozen)). These decreases are mostly the result of the training of NMT damaging the disentanglement-specific latent representation. To prevent alignment ability decline, the disentanglement module should continue to train during the NMT fine-tuning step (Sun et al., 2019).

Label Lengths

Translating short texts is well-known challenge for NMT. The bar charts in Fig. 5 show the length distribution of English ontology labels, and we can see that, with the exception of IFRS, these lengths are typically less than 7. Additionally, we computed the translation performance of the **Transformer** and **WordEmbedd+AVG** at each length. According to the BLEU score, as shown by the line charts in Fig. 5, our method performs better than Transformer for shorter ontology labels, while Transformer works better for longer labels. Despite having a little BLEU point advantage on shorter labels, our method is able to achieve superior results overall due to the high percentage of short labels. Moreover, the curve of our method is smoother than that of Transformer, and the translation performance over short texts is more stable.

Related Work

In this section, we briefly review previous studies that are related to our work. Arcan and Buitelaar (2013) employ statistical machine translation to translate ontology labels based on parallel sentences related to the ontology contents. Arcan et al. (2015); Moussallem, Soru, and Ngonga Ngomo (2019); McCrae et al. (2016) filter parallel sentences from general corpora and use them to train an ontology-specific SMT system while injecting external knowledge to switch specific domain. In contrast to earlier studies, Arcan et al. (2017) employ NMT to translate ontology.

A genre of domain adaptation method typically adds a trainable subnetwork to the NMT model. The goal is typically to improve model performance over a specific new domain. Britz, Le, and Pryzant (2017); Y. Wang, Wang, Shi, Li, and Tu (2020); Gu, Feng, and Liu (2019) add a domain classifier to identify a domain label corresponding to one of the training samples. Wu, Zhang, and Zhou (2019); Jiang, Liang, Wang, and Zhao (2020) employed an explicit multi-dimensional domain embedding instead of a classifier. Yang et al. (2019); Shi, Huang, Wang, Jian, and Tang (2019) concentrate on modeling sub-networks to represent in-domain data.

Our work is also related to research on disentangled representations. Hu, Yang, Liang, Salakhutdinov, and Xing (2017); John, Mou, Bahuleyan, and Vechtomova (2019); Cheng et al. (2020); Pergola, Gui, and He (2021) disentangle the text to the style(or sentiment) and content representations for style prediction or sentimental text generation. Bao et al. (2019); Chen, Tang, Wiseman, and Gimpel (2019) disentangle the text to the syntactic and semantic space for paraphrase generation, syntax transfer. Unlike these studies that conducted monolingual NLP research, our method focuses on multilingual settings. Additionally, some studies((B. Zhang, Xiong, Su, Duan, & Zhang, 2016b; Su et al., 2018; Sheng et al., 2020; McCarthy, Li, Gu, & Dong, 2020)) introduce variational model to enhance the translation performances. Different from these methods, we explicitly model the latent variables through variational model to enhance the translation performance.

Conclusion

We proposed architecture and multi objectives for disentangling semantic and language to improving ontology translation. The disentangled representations under cross-lingual agreement are integrated into the training and decoding phases of machine translation and induce the latent representations to capture different aspects of information through semantic matching, language discrimination, and adversarial learning. We also investigate how the disentanglement module affects translation performance. Our study demonstrates the importance of disentangled representations for addressing challenges in ontology translation.

References

Achille, A., & Soatto, S. (2018, jan). Emergence of invariance and disentanglement in deep representations. *J. Mach. Learn. Res.*, *19*(1), 1947–1980.

Arcan, M., & Buitelaar, P. (2013). Ontology label translation. In *Proceedings of the 2013 naacl hlt student research workshop* (pp. 40–46).

Arcan, M., Torregrosa, D., & Buitelaar, P. (2017). *Translating terminological expressions in knowledge bases with neural machine translation.*

Arcan, M., Turchi, M., & Buitelaar, P. (2015, July). Knowledge portability with semantic expansion of ontology labels. In *Proceedings of the 53rd annual meeting of the association for computational linguistics* (pp. 708–718). Beijing, China: Association for Computational Linguistics.

Bahdanau, D., Cho, K., & Bengio, Y. (2015, January 1). Neural machine translation by jointly learning to align and translate.. (3rd International Conference on Learning Representations, ICLR 2015)

Bao, Y., Zhou, H., Huang, S., Li, L., Mou, L., Vechtomova, O., … Chen, J. (2019, July). Generating sentences from disentangled syntactic and semantic spaces. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 6008–6019). Florence, Italy: Association for Computational Linguistics.

Bengio, Y., Courville, A., & Vincent, P. (2013). Representation learning: A review and new perspectives. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, *35*(8), 1798-1828.

Bontcheva, K. (2005). Generating tailored textual summaries from ontologies. In *European semantic web conference* (pp. 531–545).

Borst, T., & Neubert, J. (2009). Case study: Publishing stw thesaurus for economics as linked open data. *W3C Semantic Web Use Cases and Case Studies*.

Britz, D., Le, Q., & Pryzant, R. (2017, September). Effective domain mixing for neural machine translation. In *Proceedings of the second conference on machine translation* (pp. 118–126). Copenhagen, Denmark: Association for Computational Linguistics.

Buitelaar, P., Cimiano, P., Frank, A., Hartung, M., & Racioppa, S. (2008). Ontology-based information extraction and integration from heterogeneous data sources. *International Journal of Human-Computer Studies*, *66*(11), 759–788.

Cettolo, M., Girardi, C., & Federico, M. (2012). Wit3: Web inventory of transcribed and translated talks. In *Conference of european association for machine translation* (pp. 261–268).

Chen, M., Tang, Q., Wiseman, S., & Gimpel, K. (2019, June). A multi-task approach for disentangling syntax and semantics in sentence representations. 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)* (pp. 2453–2464). Minneapolis, Minnesota: Association for Computational Linguistics.

Cheng, P., Min, M. R., Shen, D., Malon, C., Zhang, Y., Li, Y., & Carin, L. (2020, July). Improving disentangled text representation learning with information-theoretic guidance. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 7530–7541). Online: Association for Computational Linguistics.

Eisele, A., & Chen, Y. (2010, May). MultiUN: A multilingual corpus from united nation documents. In *Proceedings of the seventh international conference on language resources and evaluation (LREC'10).* Valletta, Malta: European Language Resources Association (ELRA).

Fernandez, M., Lopez, V., Sabou, M., Uren, V., Vallet, D., Motta, E., & Castells, P. (2008). Semantic search meets the web. In *2008 ieee international conference on semantic computing* (pp. 253–260).

Gruber, T. R. (1995). Toward principles for the design of ontologies used for knowledge sharing. *International journal of human-computer studies*, *43*(5-6), 907–928.

Gu, S., Feng, Y., & Liu, Q. (2019, June). Improving domain adaptation translation with domain invariant and specific information. 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)* (pp. 3081–3091). Minneapolis, Minnesota: Association for Computational Linguistics.

Hermann, K. M., & Blunsom, P. (2014, June). Multilingual models for compositional distributed semantics. In *Proceedings of the 52nd annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 58–68). Baltimore, Maryland: Association for Computational Linguistics.

Hu, Z., Yang, Z., Liang, X., Salakhutdinov, R., & Xing, E. P. (2017, 06–11 Aug). Toward controlled generation of text. In D. Precup & Y. W. Teh (Eds.), *Proceedings of the 34th international conference on machine learning* (Vol. 70, pp. 1587–1596). PMLR.

Jiang, H., Liang, C., Wang, C., & Zhao, T. (2020, July). Multi-domain neural machine translation with word-level adaptive layer-wise domain mixing. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 1823–1834). Online: Association for Computational Linguistics.

John, V., Mou, L., Bahuleyan, H., & Vechtomova, O. (2019, July). Disentangled representation learning for non-parallel text style transfer. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 424–434). Florence, Italy: Association for Computational Linguistics.

Kingma, D. P., & Ba, J. (2015). Adam: A method for stochastic optimization. In *Iclr (poster)*. Retrieved from `http://arxiv.org/abs/1412.6980`

Kingma, D. P., & Welling, M. (2014). Auto-encoding variational bayes. In *Iclr.*

Koehn, P. (2005, September 13-15). Europarl: A parallel corpus for statistical machine translation. In *Proceedings of machine translation summit x: Papers* (pp. 79–86). Phuket, Thailand.

Kuang, S., Li, J., Branco, A., Luo, W., & Xiong, D. (2018, July). Attention focusing for neural machine translation by bridging source and target embeddings. In *Proceedings of the 56th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1767–1776). Melbourne, Australia: Association for Computational Linguistics.

Luong, T., Pham, H., & Manning, C. D. (2015a, June). Bilingual word representations with monolingual quality in mind. In *Proceedings of the 1st workshop on vector space modeling for natural language processing* (pp. 151–159). Denver, Colorado: Association for Computational Linguistics.

Luong, T., Pham, H., & Manning, C. D. (2015b, September). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1412–1421). Lisbon, Portugal: Association for Computational Linguistics.

McCarthy, A. D., Li, X., Gu, J., & Dong, N. (2020, July). Addressing posterior collapse with mutual information for improved variational neural machine translation. In *Proceedings of the 58th annual meeting of the association for computational linguistics* (pp. 8512–8525). Online: Association for Computational Linguistics.

McCrae, J. P., Arcan, M., Asooja, K., Gracia, J., Buitelaar, P., & Cimiano, P. (2016). Domain adaptation for ontology localization. *Journal of Web Semantics*, *36*, 23–31.

Moussallem, D., Soru, T., & Ngonga Ngomo, A.-C. (2019). Thoth: neural translation and enrichment of knowledge graphs. In *International semantic web conference* (pp. 505–522).

Papineni, K., Roukos, S., Ward, T., & Zhu, W.-J. (2002, July). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the association for computational linguistics* (pp. 311–318). Philadelphia, Pennsylvania, USA: Association for Computational Linguistics.

Pergola, G., Gui, L., & He, Y. (2021, June). A disentangled adversarial neural topic model for separating opinions from plots in user reviews. In *Proceedings of the 2021 conference of the north american chapter of the association for computational linguistics: Human language technologies* (pp. 2870–2883). Online: Association for Computational Linguistics.

Rezende, D. J., Mohamed, S., & Wierstra, D. (2014, 22–24 Jun). Stochastic backpropagation and approximate inference in deep generative models. In E. P. Xing & T. Jebara (Eds.), *Proceedings of the 31st international conference on machine learning* (Vol. 32, pp. 1278–1286). Bejing, China: PMLR.

Saunders, D., Stahlberg, F., & Byrne, B. (2019, August). UCAM biomedical translation at WMT19: Transfer learning multi-domain ensembles. In *Proceedings of the fourth conference on machine translation (volume 3: Shared task papers, day 2)* (pp. 169–174). Florence, Italy: Association for Computational Linguistics.

Sennrich, R., Haddow, B., & Birch, A. (2016, August). Neural machine translation of rare words with subword units. In *Proceedings of the 54th annual meeting of the association for computational linguistics (volume 1: Long papers)* (pp. 1715–1725). Berlin, Germany: Association for Computational Linguistics.

Sheng, X., Xu, L., Guo, J., Liu, J., Zhao, R., & Xu, Y. (2020, Apr.). Introvnmt: An introspective model for variational neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 8830-8837.

Shi, X., Huang, H., Wang, W., Jian, P., & Tang, Y.-K. (2019, November). Improving neural machine translation by achieving knowledge transfer with sentence alignment learning. In *Proceedings of the 23rd conference on computational natural language learning (conll)* (pp. 260–270). Hong Kong, China: Association for Computational Linguistics.

Stahlberg, F., Cross, J., & Stoyanov, V. (2018, October). Simple fusion: Return of the language model. In *Proceedings of the third conference on machine translation: Research papers* (pp. 204–211). Brussels, Belgium: Association for Computational Linguistics.

Steinberger, R., Ebrahim, M., Poulis, A., Carrasco-Benitez, M., Schlüter, P., Przybyszewski, M., & Gilbro, S. (2014). An overview of the european union's highly multilingual parallel corpora. *Language resources and evaluation*, *48*(4), 679–707.

Steinberger, R., Pouliquen, B., Widiger, A., Ignat, C., Erjavec, T., Tufiş, D., & Varga, D. (2006, May). The JRC-Acquis: A multilingual aligned parallel corpus with 20+ languages. In *Proceedings of the fifth international conference on language resources and evaluation (LREC'06).* Genoa, Italy: European Language Resources Association (ELRA).

Su, J., Wu, S., Xiong, D., Lu, Y., Han, X., & Zhang, B. (2018, Apr.). Variational recurrent neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, *32*(1).

Sun, H., Wang, R., Chen, K., Utiyama, M., Sumita, E., & Zhao, T. (2019, July). Unsupervised bilingual word embedding agreement for unsupervised neural machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 1235–1245). Florence, Italy: Association for Computational Linguistics.

The World Health Organization. (2011). *Icd-10 online (internet).* http://www.who.int/classifications/icd/en/. (Geneva, Switzerland: The World Health Organization) van der Maaten, L., & Hinton, G. (2008). Visualizing data using t-sne. *Journal of Machine Learning Research*, *9*(86), 2579–2605.

Van Greuning, H., Scott, D., & Terblanche, S. (2011). *International financial reporting standards: a practical guide*. World Bank Publications.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., … Polosukhin, I. (2017). Attention is all you need. In I. Guyon et al. (Eds.), *Advances in neural information processing systems* (Vol. 30). Curran Associates, Inc.

Wang, S., & Zong, C. (2017, jan). Comparison study on critical components in composition model for phrase representation. *ACM Trans. Asian Low-Resour. Lang. Inf. Process.*, *16*(3).

Wang, Y., Wang, L., Shi, S., Li, V. O., & Tu, Z. (2020, Apr.). Go from the general to the particular: Multi-domain translation with domain transformation networks. *Proceedings of the AAAI Conference on Artificial Intelligence*, *34*(05), 9233-9241.

Wu, S., Zhang, D., & Zhou, M. (2019). Effective soft-adaptation for neural machine translation. In J. Tang, M.-Y. Kan, D. Zhao, S. Li, & H. Zan (Eds.), *Natural language processing and chinese computing* (pp. 254–264). Cham: Springer International Publishing.

Yang, M., Wang, R., Chen, K., Utiyama, M., Sumita, E., Zhang, M., & Zhao, T. (2019, July). Sentence-level agreement for neural machine translation. In *Proceedings of the 57th annual meeting of the association for computational linguistics* (pp. 3076–3082). Florence, Italy: Association for Computational Linguistics.

Yazdani, M., Farahmand, M., & Henderson, J. (2015, September). Learning semantic composition to detect non-compositionality of multiword expressions. In *Proceedings of the 2015 conference on empirical methods in natural language processing* (pp. 1733–1742). Lisbon, Portugal: Association for Computational Linguistics.

Zapilko, B., Schaible, J., Mayr, P., & Mathiak, B. (2013). Thesoz: A skos representation of the thesaurus for the social sciences. *Semantic Web*, *4*(3), 257–263.

Zhang, B., Xiong, D., Su, J., Duan, H., & Zhang, M. (2016a, November). Variational neural machine translation. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 521–530). Austin, Texas: Association for Computational Linguistics.

Zhang, B., Xiong, D., Su, J., Duan, H., & Zhang, M. (2016b, November). Variational neural machine translation. In *Proceedings of the 2016 conference on empirical methods in natural language processing* (pp. 521–530). Austin, Texas: Association for Computational Linguistics.

Zhang, J., Liu, S., Li, M., Zhou, M., & Zong, C. (2014, June). Bilingually-constrained phrase embeddings for machine translation. In *Proceedings of the 52nd annual meeting of the association for computational linguistics* (pp. 111–121). Baltimore, Maryland: Association for Computational Linguistics.