INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Zero-shot Learning for Natural Language Understanding using Domain-Independent Sequential Structure and Question Types

Kugatsu Sadamitsu[1], *Yukinori Homma*[1], *Ryuichiro Higashinaka*[1], *Yoshihiro Matsuo*[1]

[1]NTT Media Intelligence Laboratories, Japan

{sadamitsu.kugatsu, homma.yukinori, higashinaka.ryuichiro, matsuo.yoshihiro}
@lab.ntt.co.jp

Abstract

Natural language understanding (NLU) is an important module of spoken dialogue systems. One of the difficulties when it comes to adapting NLU to new domains is the high cost of constructing new training data for each domain. To reduce this cost, we propose a zero-shot learning of NLU that takes into account the sequential structures of sentences together with general question types across different domains. Experimental results show that our methods achieve higher accuracy than baseline methods in two completely different domains (insurance and sightseeing).

Index Terms: natural language understanding, zero-shot learning, question answering

1. Introduction

Research on natural language understanding (NLU), a process that converts sentences into queries for knowledge bases (KBs), has been gaining popularity as spoken dialogue systems become increasingly used by the general public [1, 2, 3, 4, 5]. However, when we adapt NLU to new domains, it is quite expensive to construct training data with the mappings between words in a question sentence and an entry of KB. To reduce the cost, several transfer learning and zero-shot learning methods have been proposed [6, 7, 8, 9, 10, 11]. These methods need little or no training data about new domains. In zero-shot learning, word similarity scores based on word-embedding features about words in a question sentence and an entry in KBs are utilized. While this zero-shot learning approach is effective in similar domains, it is difficult to maintain high performance when we apply it to completely different domains because of the lack of information about the target domain (e.g., context and word similarity information).

In this paper, we propose a new zero-shot learning of NLU that utilizes common knowledge across domains, especially the sequential structure of question sentences and the external knowledge of general question type (QT) for covering the lack of information about the target domain. As the external knowledge, we utilize the relationship between the QT corresponding to the entry in a KB and the QT of the question sentence estimated by our general classifier trained from large supervised corpora. As far as we know, there have been no prior studies on adapting general question types for zero-shot learning for NLU.

We performed experiments to evaluate our proposed methods in two completely different domains ("sightseeing" and "insurance") and found that they were more accurate than baseline methods.

2. Related Work

NLU tasks have conventionally been studied using a supervised approach [1, 2, 3, 4, 5] for known domains. To augment supervised approaches, transfer learning has been proposed in order to improve the NLU accuracy of target domains with data from target and other domains [6, 7, 10, 11]. In conventional transfer learning approaches, Li et al. and Kim et al. proposed label mapping from the source domain to the target domain directly [6, 7]. However, they had difficulties providing adequate coverage of the mapping of entries when a large difference exists between the entries of the domains. In contrast, because we introduce general question types, our method can be adapted to new domain.

Zero-shot learning approaches for NLU have recently been proposed. Generally speaking, conventional zero-shot learning is achieved by utilizing the clue of similarity in the same vector spaces of source and target domains [12, 13, 14]. Yazdani and Henderson applied zero-shot learning to NLU between similar domains using word embedding to construct the same vector space for question sentences and the entries of KBs [8]. They learn a binary classifier in the vector space in the source domain to adapt to target domains. However, it is difficult for their method to achieve high performance in completely different domains.

3. Task Settings

Our ultimate objective is to be able to query any KBs by input question sentences in natural language. We take a zero-shot learning approach so that we can avoid the high cost associated with constructing training data for each domain. To this end, we need to extract essential queries from an input question sentence and map them to KB entries in an unknown domain. A recent task in zero-shot learning for NLU, especially in the context of dialogue processing, is the extraction of concepts or arguments of dialogue action (dialogue act) from an input sentence [8]. For example, from the input sentence "*I would like Chinese food.*", the system extracts "*inform (food = Chinese)*", where "*inform*" is a dialogue action and the items inside the brackets are the concepts associated with the dialogue action.

Our task setting is similar. Our KBs are assumed to be constructed of Resource Description Framework (RDF) triples, namely, "subject, predicate, object" (hereinafter *subj*, *pred*, and *obj*) and their components are called triple entries g whose values are written in natural language (e.g., g ="*pred:ACCESS*" and g ="*obj:¥1,000*" in Fig. 1). We call the type of these triple entries "triple symbol" $c \in \{subj, pred, obj\}$. Our purpose is to extract each triple entry from the input question sentence.

Zero-shot NLU models are trained by data consisting of the KB and question sentences written in natural language anno-

http://dx.doi.org/10.21437/Interspeech.2017-638

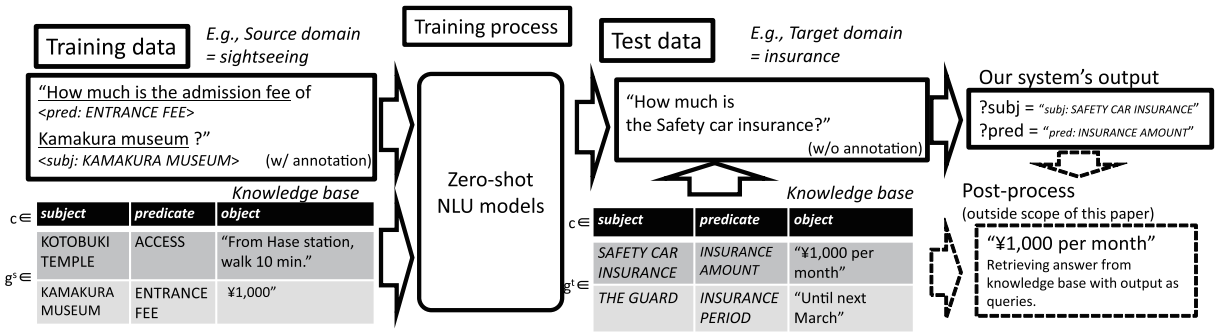

Figure 1: *Overview of zero-shot learning task.*

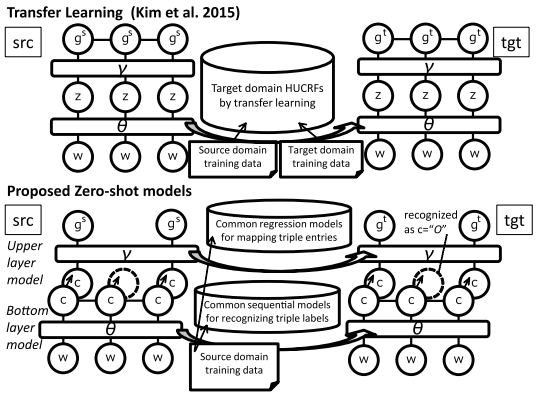

Figure 2: *Comparison of transfer learning (upper) and proposed method (lower).*

tated with KB queries about a source domain (e.g., some words in a sentence are underlined and annotated as "*<pred: ENTRANCE FEE>*" in Fig. 1). In the test step, the input is a KB and a question sentence without annotation about the target domain (e.g., "insurance" in Fig. 1), and the output is a query for the KB in the target domain.

4. Proposed Method

The zero-shot learning methods proposed thus far have not used domain-independent knowledge [8, 9]. We propose a new method for zero-shot learning that can take into account domain-independent knowledge, especially question sentence structure information and general question types (QTs) information. Our method comprises a new model based on sequential models (Sec. 4.1) and new features pertaining QT (Sec. 4.2).

4.1. Proposed Models

In this section, we propose a model for zero-shot learning that can take into account the question sentence structure similar to conventional transfer learning models[7].

4.1.1. Sequential Transfer Learning

Kim et al. [7] proposed a transfer learning that utilizes hidden unit CRFs (HUCRFs) [15] for treating sequential information. HUCRFs include the hidden middle layer $z \in \{0, 1\}$ between the input and output layers. The graphical expression is shown in the upper part of Fig. 2 and the probabilities of the output sequence of triple entries $\mathbf{g}$ given input words $\mathbf{w}$ are formulated as $p(\mathbf{g}|\mathbf{w}) = \sum_{\mathbf{z}\in\{0,1\}^n} p(\mathbf{g}, \mathbf{z}|\mathbf{w})$, where $\mathbf{z} \in \{0, 1\}^n$ are hidden variables, $p(\mathbf{g}, \mathbf{z}|\mathbf{w}) \propto \exp(\theta^T \mathbf{\Phi}(\mathbf{w}, \mathbf{z}) + \gamma^T \mathbf{\Psi}(\mathbf{z}, \mathbf{g}))$, n is the total number of words, and $\theta \in \mathbb{R}^d, \gamma \in \mathbb{R}^{d'}$ are the parameters of the lower and upper layer models in Fig. 2, respectively, and $\Phi(\mathbf{w}, \mathbf{z}) \in \mathbb{R}^d, \Psi(\mathbf{z}, \mathbf{g}) \in \mathbb{R}^{d'}$ are their respective feature functions.

Kim et al. [7] made an assumption that the parameters of the lower layer θ tend to be similar across domains and proposed a method for transferring the parameters of the lower layer θ of HUCRF to the target domain, as shown in Fig. 2. Their method is a reasonable approach in the transfer learning task because the parameter γ can be estimated by utilizing target domain training data. However, in zero-shot learning tasks, it is hard to adapt the parameters of γ without target domain training data.

4.1.2. Sequential Zero-shot Learning Models

Because the conventional sequential model based transfer learning needs training data in the target domain [7], it cannot be adapted to zero-shot learning straightforwardly. We take the concept of the transfer learning and adapt it to zero-shot learning models.

The lower side of Fig. 2 shows our proposed model, where w is a word and g^s, g^t are triple entries, with s, t indicating source and target domains. Our model is constructed using two cascade-connected layers. First, the lower layer (θ) estimates the highest likelihood sequence of triple symbols $\mathbf{c}$, which is common between domains, and then the upper layer (γ) estimates the mapping to triple entries g^s or g^t from all of the triple entry candidates G_c^s or G_c^t whose triple symbols are corresponding with triple symbol c.

We extend the conventional transfer models in two ways. First, we extend the middle layer variable $z \in \{0, 1\}$ to triple symbol $c \in \{subj, pred, obj\}$ with BIO labels. Extending the variable of the middle layer makes it possible to propagate rich information from input w. For example, the typical context "What is the *<pred>* of *<subj>*?" indicate the typical positions of triple symbols, *subj* and *pred*.

To construct the lower layer, we utilize CRFs for labeling triple symbol c. In the training step, the annotation about triple entry g in the source domain has been abstracted to triple symbol c. For example, the annotated label g ="*pred:ENTRANCE FEE*" is replaced by c =*pred*. This will enable the trained CRFs

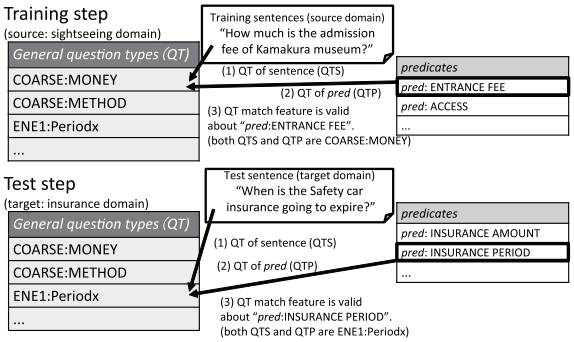

Figure 3: *The utilization of general question types*

to be adapted to unknown domains. The features for the CRFs are word surface forms, POS, ngrams, and word-embedding features using CBOW [16]. For parameter estimation, we utilize minimum classification error criteria [17].

Second, we make it possible to transfer the parameters for triple entry mapping (γ), not only θ. The upper layer maps estimated triple symbols c to triple entries g by using domain-independent multiple similarity features in the logistic regression models [18].

In the training step, for constructing training data, we give 1 for correct mapping to g^s and 0 for incorrect mapping $\bar{g}^s$, which are randomly selected triple entry candidates in G^s_c. In the adaptation step, the regression model provides a value for each candidate of triple entry g^t and we select the entry with the highest value as the estimation result; that is

$$g^t_i = \arg \max_{g^t \in G^t_{c_i}} p(\gamma^T \cdot \mathbf{\Psi}(g^t, c_i, w_i)),$$

where i means a word position, γ indicates the weight of each feature in a regression model, and Ψ is the feature function described in the next section.

4.2. General Features for Triple Entry Mapping

In the process of mapping triple symbols c to triple entries g in the regression model (based on γ), the design of general features ($\mathbf{\Psi}$) across domains is very important. We introduce effective features proposed in previous studies as well as our own new features pertaining to general question types.

4.2.1. Basic Features

In accordance with the previous studies, we utilize the basic features between words in a chunk $\mathbf{w}'$ and a triple entry g; that is, we use surface similarities (word correspondence ratio, character overlaps, and edit distances) and semantic similarities based on word embeddings [8, 9]. For the word-embedding features, we use cosine similarities from domain-dependent and domain-independent CBOW models [16][1]. Here, the domain-dependent word-embedding models are trained using 186K WEB sentences retrieved by queries using both target and source domain triple entries (e.g., "entrance fee" and "insurance amount" in Fig. 1). The domain-independent word-embedding models are trained by 12M Wikipedia sentences. The number of dimensions is 100 in both models.

[1] https://code.google.com/p/word2vec

4.2.2. Features of General Question Types

Although the basic features are effective, input words and triple entries might not be similar in the measurements of the surface and word-embedding. In particular, *pred*s appear more often in surfaces different from each KB entry; for example, the word "expire" is not similar to "*pred:INSURANCE PERIOD*" from the perspective of a surface and word-embedding.

To tackle this issue, we propose utilizing general question types (QTs) as new features. This is based on our assumption that, even if the domain changes greatly, the basic intention of the question would likely be similar. Our approach hinges on our observation that the question types of a question sentence (QTS) and of a *pred* (QTP) tend to correspond. For example, in Fig. 3, "expire" in a test sentence is often used for asking about "period" (QTS) and "*pred:INSURANCE PERIOD*" is also a type of "period" (QTP). We utilize these correspondences as new features.

We describe how to utilize QT as features in detail. We utilize two types of QT based on Higashinaka et al.'s work [19]. The first type is "QT-coarse", which is a coarse classification with 23 classes including non-factoid question types [20]. The second type is a finer granularity classification for factoid questions based on Sekine's extended named entity (ENE) definition [21]. The ENE definition has three layers; we utilize the first (QT-ENE1: 28 types) and second (QT-ENE2: 87 types) layers, abandoning the third because it is too fine-grained and might cause sparseness problems.

To utilize QT information, first, we estimate the QTS of an input question sentence by a QTS classifier using training data prepared in advance. Second, we retrieve the QTP of the candidate *pred* g^t, and finally, we utilize the correspondence between QTS and QTP as features for entry mapping. The detailed steps are shown below (each item number corresponds to the number in Fig. 3).

1. The input question sentence is classified into top N QT labels on the basis of the QTS classifier with their probabilities ($N = 3$). We use a logistic regression classifier [18] for estimating QTS with standard features including surfaces and ngrams [2] .
2. For determining QTP, we map the *pred*s in G_{pred} to QT by measuring the distance of word-embeddings [16] between the name of QT labels and *pred* entries with human confirmation.
3. We utilize features expressing the correspondence between QTS and QTP in two ways. The first is a match feature, "QT-match", that indicates whether QTS and QTP correspond or not. The second is a pairwise feature, "QT-pair", which is simply a pair of QTS and QTP used as a feature. The QT-match feature is robust for an unknown domain because it only considers a match; it does not consider whether the QT types appeared in the training data or not. The more the QT-pair feature is used, the more precise the detection becomes compared to the QT-match feature. Each value of the feature is represented by the probability of the estimated QTS.

For an example question sentence (shown in Fig. 3), "*When is the Safety car insurance going to expire?*", the QTS is estimated

[2] The training data for the QTS classifier is designed to cover every QT type with 63,843 and 56,629 sentences for "QT-COARSE" and "QT-ENE1/2", respectively. The classification accuracy by two-fold cross validation is 86.94%, 86.01%, and 78.61% in "QT-coarse", "QT-ENE1", and "QT-ENE2", respectively.

Table 1: *Comparison of accuracy. Underlined F values are the best results in each domain. $*$ and † indicate significant difference ($p < .01$) compared to "Baseline 1" and "w/o QT".*

	src:insurance to tgt:sightseeing			src:sightseeing to tgt:insurance		
	prec	rec	F	prec	rec	F
Baseline 1	34.3	26.2	29.7	21.5	58.6	31.4
Baseline 2	22.7	13.1	16.6	39.5	50.0	44.2
Proposed model (w/o QT)	46.8	24.1	31.9	42.5	53.5	47.4*
+QT match	62.2	33.4	43.5*†	42.7	54.1	47.7*
+QT match + QT pair	62.9	34.2	<u>44.3*†</u>	43.9	55.8	<u>49.1*†</u>
Supervised in-domain model (upper bound)	src:sightseeing to tgt:sightseeing			src:insurance to tgt:insurance		
	55.5	54.9	55.2	58.9	54.3	56.5

as "ENE1:Periodx" by the QTS classifier (1). One of the triple entry candidates "*pred:INSURANCE PERIOD*" is mapped to QT "ENE1:Periodx" for QTP (2), and then the "QT-match" feature and the "QT-pair" feature pertaining to "ENE1:Periodx-ENE1:Periodx" become valid (3).

5. Experiments

5.1. Experimental settings

We examine the effectiveness of our model in two completely different domains ("sightseeing" and "insurance"). We made annotated question sentences for training and evaluation in each domain. The amount of data is 937 sentences including 825 *subj*, 868 *pred*, and 371 *obj* for "sightseeing" and 443 sentences including 289 *subj*, 393 *pred*, and 368 *obj* for "insurance". For KBs, we use 92 types of subjects and 44 types of predicates for the "sightseeing" schema derived from Wikipedia's infobox and 16 types of subjects and 29 types of predicates for the "insurance" schema derived from a pamphlet on insurance products. Both text data and KBs are in Japanese. We evaluate our methods by precision, recall, and F values.

For the evaluation of our proposed methods, we examine three conditions related to QT: proposed sequential zero-shot models without QT (w/o QT), with QT-match features (+QT match), and with both QT-match and QT-pair features (+QT match + QT pair). The *pred*s that are dependent on QT are evaluated under every condition. Note that the results for *subj* and *obj* are mostly independent of QT.

We prepared two baselines. The first one deals with each word without context as in previous work [8, 9]. We include this baseline for the verification of our proposed model utilizing sequential information. The second baseline uses direct entry mapping between the source and target entries in each domain as in [6, 7]. The direct mapping of entities between source and target domains are manually annotated. Some entities in the target domain do not have mapping that has not appeared in source domain. We include this baseline for the comparison between our proposed method utilizing external knowledge (i.e., QT) and the existing direct mapping method.

Furthermore, for confirming the upper bound accuracy, we examined in-domain supervised learning by 10-fold cross validation using the same testset. Note that, with this procedure, the amount of training data is reduced to $9/10$.

5.2. Experimental results and analysis

The results in Table 1 demonstrate that the proposed model without QT improved the accuracy more than Baseline 1, particularly in terms of precision (34.3 → 46.8 and 21.5 → 42.5). We also found that the QT features with external knowledge further improved the performance (in F value, $31.9 \rightarrow 44.3$ and $47.4 \rightarrow 49.1$) in contrast to Baseline 2 which was worse than w/o QT. The last row in Table 1 shows that although there is still a gap between in-domain supervised and zero-shot learning, the proposed methods reduced this gap (they achieved about 80.3% and 86.9% the performance of the in-domain supervised method). The other results for triple symbols *subj* and *obj* are shown in Table 2 in F values. From these results, the accuracy of *subj* in the insurance to sightseeing adaptation became worse than the baseline, as some of the *subj*s were mistakenly recognized as *pred*; however, total accuracy was improved ($53.2 \rightarrow 55.5$ and $46.4 \rightarrow 58.6$).

Table 2: *The results of each triple symbol in F values.*

	src:insurance to tgt:sightseeing				src:sightseeing to tgt:insurance			
	all	*pred*	*subj*	*obj*	all	*pred*	*subj*	*obj*
Proposed	55.5	44.3	69.0	32.1	58.6	49.1	71.8	60.5
Baseline 1	53.2	29.7	76.3	31.3	46.4	31.4	67.3	61.7

An improved and a worsened examples are:

- (Improved) "How much per hour is the car park?": The baseline 1 was misrecognized as asking about "*pred:EXISTENCE OF PARKING*" because of high similarity scores about surfaces.
- (Worsened) "Where does the name of Jufukuji Temple come from?": The proposed method misrecognized as asking about "*pred:LOCATION*" because of the word "where". However, there are few such examples among our results.

Finally, we mention some feature weights trained in the "sightseeing" domain. Note that all feature values have normalized values between 0 and 1. For the basic features, the weights were 1.67 for surface and 2.37 for word-embedding. For the QT features, the weights were $1.28, 0.61$, and 0.71 for the QT-match of QT-Coarse, QT-ENE1, and QT-ENE2, respectively. These training results demonstrate the effectiveness of using the QT features.

6. Conclusion

We proposed a zero-shot learning of NLU that can deal with sequential structures and general question types across different domains. Experimental results demonstrate that both the proposed features and the model are effective.

In future work, because the training with a single domain might depend on that domain even with our general features, we will apply our method to more than two domains. This will lead to the discovery of common features across multiple domains.

7. References

[1] K. Yao, B. Peng, G. Zweig, D. Yu, X. Li, and F. Gao, "Reccurent Conditional Random Field for Language Understanding," in *Proceedings of the Acoustics, Speech and Signal Processing*, 2014, pp. 4077–4081.

[2] K. Yao, G. Zweig, M.-y. Hwang, Y. Shi, and D. Yu, "Recurrent Neural Networks for Language Understanding," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2013, pp. 104–108.

[3] R. J. Kate and R. J. Mooney, "Using string-kernels for learning semantic parsers," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, 2006, pp. 913–920.

[4] S. Pradhan, W. Ward, K. Hacioglu, and J. H. Martin., "Shallow semantic parsing using support vector machines," in *Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics*, 2004, pp. 233–240.

[5] R. Schwartz, S. Miller, D. Stallard, and J. Makhoul, "Language understanding using hidden understanding models," in *Proceedings of Fourth International Conference on Spoken Language Processing*, vol. 2, 1996, pp. 997–1000.

[6] X. Li, Y. Y. Wang, and G. Tur, "Multi-task learning for spoken language understanding with shared slots," in *Proceedings of the Annual Conference of the International Speech Communication Association*, 2011, pp. 701–704.

[7] Y.-B. Kim, K. Stratos, R. Sarikaya, and M. Jeong, "New Transfer Learning Techniques for Disparate Label Sets," in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing*, 2015, pp. 473–482.

[8] M. Yazdani and J. Henderson, "A Model of Zero-Shot Learning of Spoken Language Understanding," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, 2015, pp. 244–249.

[9] E. Ferreira, B. Jabaian, and F. Lefevre, "Online Adaptive Zero-Shot Learning Spoken Language Understanding Using Word-Embedding," in *Proceedings of the Acoustics, Speech and Signal Processing*, 2015, pp. 5321–5325.

[10] G. Tur, "Multitask learning for spoken language understanding," in *Proceedings of the Acoustics, Speech and Signal Processing*, 2006, pp. 585–588.

[11] M. Jeong and G. Geunbae Lee, "Multi-domain spoken language understanding with transfer learning," *Speech Communication*, vol. 51, no. 5, pp. 412–424, 2009.

[12] M. Palatucci, G. E. Hinton, D. Pomerleau, and T. M. Mitchell, "Zero-Shot Learning with Semantic Output Codes," in *Proceedings of the Advances in Neural Information Processing Systems*, 2009, pp. 1410–1418.

[13] J. Weston, S. Bengio, and N. Usunier, "Wsabie: Scaling up to large vocabulary image annotation," in *Proceedings of the 22nd International Joint Conference on Artificial Intelligence*, 2011, pp. 2764–2770.

[14] J. Weston, S. Bengio, and N. Usunier, "Large scale image annotation: Learning to rank with joint wordimage embeddings," in *European Conference on Machine Learning*, 2010, pp. 21–35.

[15] L. Maaten, M. Welling, and L. K. Saul, "Hidden-unit conditional random fields," in *Proceedings of the International Conference on Artificial Intelligence and Statistics*, 2011, pp. 479–488.

[16] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," *arXiv preprint arXiv:1301.3781*, 2013.

[17] J. Suzuki, E. McDermott, and H. Isozaki, "Training Conditional Random Fields with Multivariate Evaluation Measures," in *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*, 2006, pp. 217–224.

[18] R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin, "LIBLINEAR: A library for large linear classification," *Journal of Machine Learning Research*, vol. 9, pp. 1871–1874, 2008.

[19] R. Higashinaka, K. Sadamitsu, K. Saito, and N. Kobayashi, "Question answering technology for pinpointing answers to a wide range of questions," *NTT Technical Review*, vol. 11, no. 7, 2013.

[20] M. Nagata, K. Saito, and Y. Matsuo, "Japanese natural sentence search system Web Answers (in Japanese) ," in *Proceedings of the Annual Conference of the 12th Annual Meeting of the Association for Language Processing*, 2006, pp. 320–323.

[21] S. Sekine, "Extended named entity ontology with attribute information," in *Proceedings of the 6th International Language Resources and Evaluation*, 2008.