Proceedings of the 12th Conference on Language Resources and Evaluation (LREC 2020), pages 2796–2802
Marseille, 11–16 May 2020

Building a Task-oriented Dialog System for languages with no training data: the Case for Basque

Maddalen López de Lacalle, Xabier Saralegi, Iñaki San Vicente
Elhuyar Foundation
Osinalde Industrialdea 3, 20170 Usurbil, Spain
{m.lopezdelacalle, x.saralegi, i.sanvicente}@elhuyar.eus

Abstract

This paper presents an approach for developing a task-oriented dialog system for less-resourced languages in scenarios where no training data is available. Both intent classification and slot filling are tackled. We project the existing annotations in rich-resource languages by means of Neural Machine Translation (NMT) and posterior word alignments. We then compare training on the projected monolingual data with direct model transfer alternatives. Intent Classifiers and slot filling sequence taggers are implemented using a BiLSTM architecture or by fine-tuning BERT transformer models. Models learnt exclusively from Basque projected data provide better accuracies for slot filling. Combining Basque projected train data with rich-resource languages data outperforms consistently models trained solely on projected data for intent classification. At any rate, we achieve competitive performance in both tasks, with accuracies of 81% for intent classification and 77% for slot filling.

Keywords: Dialog Systems, Neural language representation models, Less-resourced languages

1. Introduction

Task-oriented dialog systems are one of the most trending topics in today's Natural Language Processing. Such systems are focused on helping users with specific tasks such as booking a ticket or a restaurant, scheduling a reminder for a call, etc.

Task-oriented dialog systems understand what users mean when they speak or write an utterance (the request of the user) by identifying users intents and the corresponding slots. Intents are what the user wants, and slots are the key entities corresponding to the intent. Let's take a look at example 1. The user's intent is to set a new reminder. Intents are given a name such as *"set reminder"* in this case. Intent-arguments or slots modify the intent, which in this case are *"to take out the garbage"* and *"on Thursday"*. Intent-arguments are given a name, such as *"to do"* and *"datetime"* in our example.

Example 1. *Remind me [to take out the garbage] [on Thursday]*

The main tasks involved in a task-oriented dialog system are intent classification and slot filling, which can be stated as sentence classification and sequence labeling problems respectively. These two tasks can be learnt in a joint fashion or as two separate tasks. Most successful approaches proposed for both strategies rely on deep learning algorithms (Xu and Sarikaya, 2013; Mesnil et al., 2013; Liu and Lane, 2016; Goo et al., 2018), and hence, they require large training datasets. Unfortunately, building labeled dialog datasets involves a great manual effort, rendering it unfeasible for non major languages. Scenarios where task-oriented dialog systems must deal with many languages can experiment a similar problem, because the cost for building training datasets multiplies.

In this work we deal with the problem of building a task-oriented dialog system for a language without any native training data available. Our aim is to find out whether it is possible to transfer knowledge included in datasets from other languages, or even systems trained for other languages. We focus on Basque as a case study, a less-resourced language with rich morphology and syntax different from any major language.

For performing the cross-lingual transfer learning we compare two approaches:

1. Projection of training data: Translating training data from rich-resource languages and project the annotation by word alignment. Intent classifier and slot labeler for Basque are learnt from projected training data.

2. Direct model transfer: Training intent classifiers and slot labelers from training data in rich-resource languages by using multilingual language models.

These two main approaches have been used previously in sentence classification and sequence labeling type tasks such as NERC (Mayhew et al., 2017), POS (Yarowsky et al., 2001) and even slot filling (Schuster et al., 2019). The main contributions of this paper regarding those works are the following: (i) We propose a method which combines Transformer based NMT (Vaswani et al., 2017) and word alignment (Dyer et al., 2013) for projecting slot annotations to Basque from a rich-resource language; (ii) We provide a comparison of state of the art algorithms for dealing with sentence classification and sequence labeling applied to intent classification and slot filling, paying special attention to multilingual language models such as BERT (Devlin et al., 2019), that allow for direct model transfer strategies; (iii) We present pioneer work for Basque on these tasks. We have built the first training and test datasets in this language, and made them publicly available[1] for future benchmarking.

[1] https://hizkuntzateknologiak.elhuyar.eus/assets/files/fmtodelh.tgz

From here on the paper is structured as follows. Section 2. describes the data used in the experiments. Next section presents the method implemented to project Spanish data to Basque. Sections 4. and 5. give details on the algorithms applied to tackle intent classification and slot filling tasks, respectively. Results obtained in the experiments are described in section 6. Previous work on the field is addressed in section 7. and we end the paper drawing conclusions on the experiments carried out.

2. Dataset

There are several datasets related to the tasks of intent classification and detection and classification of intent-arguments or slots. Among others, we can find datasets related to airline travels (Tur et al., 2010), restaurant booking (Henderson et al., 2014) and multi-domain (SNIPS[2]). These datasets are made up of sets of utterances that include annotations related to the type of intention and the various associated arguments.

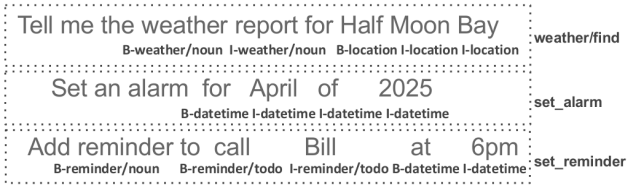

Figure 1: Examples of utterances included in *FMTOD*. Slots in bold below the utterances and intent on the right.

Unfortunately, most of those datasets are in English, which complicates experimentation with other languages. For this work, which aims precisely at the study of different techniques for the generation of intent and arguments classifiers for the case of languages without training datasets, we have chosen to adapt the evaluation data of an existing dataset to Basque, specifically we carry out the experiments using the Facebook Multilingual Task Oriented Dataset (*FMTOD*) (Schuster et al., 2019).

We have selected the *FMTOD* for two reasons: a) it includes the most common domains of an assistant, and b) it contains utterances in Spanish which allow us to apply a Spanish to Basque machine translation system to project the Spanish dataset into Basque.

The *FMTOD* contains manually generated and annotated utterances for three languages: English (*FMTODen*), Spanish (*FMTODes*), and Thai (*FMTODth*). Figure 1 shows some English examples. They are grouped into three domains (alarm, reminder and weather) and classified according to 12 types of intentions that include up to 11 types of arguments (See statistics in Table 2). As the cost of manually translating the full dataset to Basque is out of reach with our resources, we settled for preparing a reliable test dataset. We randomly selected 100 utterances in English for each type of intention. Next, English statements (1,348) were translated into Basque by a professional translator, and finally, a native speaker annotated the arguments. We will refer to this dataset as *FMTODeu*.

We have already mentioned that we took *FMTODes* as starting point for projecting the training data into Basque. However, we did not use the original dataset. An initial review of the Spanish dataset, revealed that a significant number of translations included remarkable spelling, morphological, syntactic or lexical failures. For its correction we reviewed the development, train and test sets and proceeded to make the corresponding corrections of both the texts and the annotations. The utterances that included a high number of failures were directly discarded to lighten the manual effort. For the same reason the entire test was not reviewed. We refer to this corrected version of *FMTODes* as *FMTODesv2*. Table 1 offers the statistics of the final datasets we used in the experiments.

Dataset	*FMTODen*	*FMTODesv2*	*FMTODeu*
Train	30,521	3,418	3,418 (projected)
Dev	4,181	1,900	1,900 (projected)
Test	8,621	1,348	1,086 (manual)

Table 1: Statistics of datasets used in the experiments.

Domain	Intent type	Slot type
Alarm	6	2
Reminder	3	6
Weather	3	5
Total	12	11

Table 2: Domain, intent types and slot types included in *FMTOD*.

3. Annotation projection

Projecting slot filling data presents an additional challenge, because in addition to translating the examples, the offsets of the slots must be mapped to the target language. (Yarowsky et al., 2001) proposed a general solution to create training data for sequence taggers in a language from existing datasets in other languages. This approach involves the use of parallel corpora, or an automatic translator, and the projection of annotations according to a word level alignment. (Yarowsky et al., 2001) initially used the approach to generate taggers (part-of-speech, baseNP, NERC, and inflectional morphological analysers) for several languages starting from English data. Since then variants of this approach have been proposed by different authors for various tasks such as NER (Kim et al., 2012; Ni and Florian, 2016; Mayhew et al., 2017), POS (Duong et al., 2014) and even the task at hand, i.e., Slot Filling (Schuster et al., 2019).

Our projection process consists of three steps (See example in Figure 2):

[2] https://github.com/snipsco/nlu-benchmark/tree/master/2017-06-custom-intent-engines

1. Utterances of train and dev sets of *FMTODesv2* are translated by means of an Spanish to Basque Transformed based Neural Machine Translation system (Vaswani et al., 2017) trained on a parallel corpus of 10 million segments. This system achieved a BLUE of 23.02 over a test of 5,000 segments.

2. Word level alignment is performed between the utterance in Spanish and its translation in Basque. For that aim, we use two alignment models trained using the IBM Model 2 variant proposed by (Dyer et al., 2013). We train a first alignment model from source to target language and another one from target to source language. These models are trained from a parallel corpus of 6 million segments, and they are applied to each pair of utterances (Spanish source and Basque translation) using the heuristic *grow-diag-final-and* for symmetry.

3. Projection of slot labels according to the alignment between words, and automatic repair of inconsistencies in the BIO encoding.

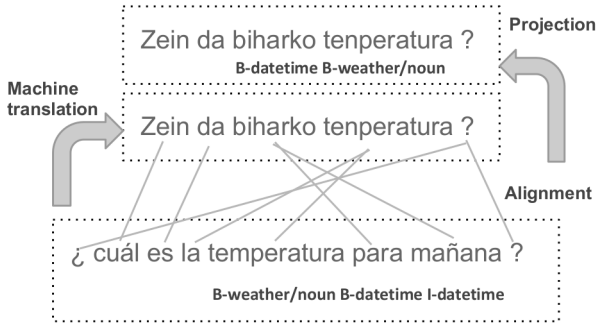

Figure 2: Example of projection of slot annotation from Spanish to Basque.

4. Intent classification

The correct identification and classification of users' intent is decisive in a dialog system. Usually, users tend to express their intention with short and lack of context utterances so traditional rule-based or machine-learning methods are not good enough. Such methods depend on the grammatical accuracy of the sentences and they can identify explicit aspects but fail to extract implicit aspects (Hashemi et al., 2016).

Nowadays neural network-based architectures dominate document classification tasks, and currently they represent the state of the art for intent classification as well. (Adhikari et al., 2019) propose a properly-regularized single-layer BiLSTM. (Vaswani et al., 2017) presents a Transformer architecture where a fully-connected layer is added over the final hidden state, and (Devlin et al., 2019) implement a multi-layer bidirectional Transformer encoder.

In this work we focus the identification and classification of the intents as a sentence classification task. We have compared two different approaches which do not need native training dataset in the target language. We use two different cross-lingual transfer methods: (1) automatically translating the Spanish training data with a NMT system to Basque and projecting the annotations; and (2) transferring directly the models trained on the native datasets. We also combine the datasets in different languages to train multilingual classifiers and compare them to classifiers that are trained exclusively on the automatically translated dataset.

Once we have generated the training data, we have to fine-tune our language models and train classifiers on the downstream task. For that aim we have selected two architectures:

BERT fine-tuning (Devlin et al., 2019): introduced BERT (Bidirectional Encoder Representations from Transformers), transformer-based language model trained on enormous quantities of texts which has achieved state of the art results on several NLP tasks. There is no BERT monolingual model for Basque, but multilingual BERT (mBERT), does include Basque among its languages. mBERT is a single language model pre-trained from corpora in 104 languages. Hence it is an adequate choice for our direct model transfer strategy. Moreover, mBERT can be easily fine-tuned on downstream tasks, by training for a few epochs on the data of the specific task. In our experiments we use the same fine-tuning strategy as the original paper, feeding the output [CLS] token representation to an output layer for classification.

Flair (Akbik et al., 2018) is a deep learning system achieving state of the art results in several tasks. Its major strengths are their own character level contextual embeddings, and the ability to apply straightforwardly a large number of embedding types to train classifiers in downstream tasks. In addition to this, Flair provides native embeddings for Basque which allows us comparison with the multilingual models.

For intent classification, Flair official embeddings (we stack forward and backward models in all our settings) are fed into a BiLSTM[3] to produce a document level embedding which is then used in a linear layer to make the class prediction. We did not try further embedding combination, although Flair developers recommend to stack their own Flair embeddings with additional static embeddings such as FastText.

5. Slot filling

The second task we deal with is the identification and classification of the so-called slots or intent-arguments. As mentioned before, we address this task as a sequence labeling problem. As for the intent classification task, we compare the same two different cross-lingual transfer methods: (1) translation and projection of the training data and (2) transferring directly the models trained on the native dataset.

Current state-of-the-art approaches for sequence labeling typically use the LSTM variant of bidirectional recurrent neural networks (BiLSTMs), and a subsequent Conditional Random Field (CRF) decoding layer (Huang et al., 2015; Ma and Hovy, 2016). Thus, we study implementing the

[3]A single layer of size 128 was used, with word reprojection and a dropout value of 0.3068.

sequence labeler using a BiLSTM-CRF architecture with contextualized language models. Similar to section 4., we train Flair (Akbik et al., 2018) contextual word embeddings in the slot-filling task. In this case stacked forward and backward embeddings are feed through a BiLSTM RNN connected to a CRF decoding layer.
Experiments regarding direct transfer model are carried out again making use of the BERT fine-tuning approach explained in the previous section. The only difference is that the output tokens are feed into a CRF layer.

6. Results

This section presents the results of the experiments carried out for intent classification and slot filling. For both tasks, firstly we present a comparison of the various systems trained solely on monolingual data. Secondly, we provide comparison between systems relying on projected training data, direct model transfer strategies and finally combination of multilingual training data.
All the tables in the following subsections present the averaged result of 5 randomly initialized runs. Slot filling was evaluated using official CoNLL evaluation script[4].

6.1. Intent classification

Flair related experiments were conducted using the Flair text classifier with the same hyperparameter setting[5].
BERT fine-tuning was done with a learning rate of 2e-5 and a batch size of 16. The number of epochs to fine-tune was selected based on the results from monolingual experiments (Table 3), which were optimized on the dev set, to the point were standard deviation of micro F1-score between the five runs was lower than 1. For all systems including Basque training data the number of epochs was 15, the system trained on solely English data was fine-tuned for 3 epochs and the system trained solely on Spanish data was fine-tuned for 10 epochs. As all of those classifiers rely on the multilingual BERT model, our intuition is that the system requires a longer fine-tuning when the target language is less represented within the BERT model.

	EN	ES	EU
BERT monolingual	99.23	-	-
mBERT fine-tuning	99.23	98.03	78.56
Flair Embeddings	99.05	96.95	76.21
Previous results			
(Schuster et al., 2019)	99.11	97.61	-

Table 3: Micro F1 results for Intent classification.

Table 3 shows the results for the monolingual intent classification task. The first thing we notice is that mBERT performs on par with its monolingual counterpart in English. Secondly, mBERT fine-tuned models slightly outperform Flair monolingual models for all languages. Moreover, fine-tuned mBERT model also outperforms the results in (Schuster et al., 2019) for English and Spanish. Lastly, it is remarkable that mBERT achieves the best results for Basque, even if the presence of the Basque language is very small in mBERT.

	micro F1	Macro F1
Monolingual		
EU (Projected)	78.56	68.31
Direct model transfer (cross-lingual)		
EN_{train}/EU_{test}	31.48	18.84
ES_{train}/EU_{test}	29.44	18.95
Multilingual training data		
$ES + EU_{train}/EU_{test}$	79.52	71.48
$EN + EU_{train}/EU_{test}$	81.01	75.05
$ES + EN + EU_{train}/EU_{test}$	**81.56**	**75.62**

Table 4: Intent classification results for direct transfer model experiments.

Table 4 shows the results achieved by the different training strategies. All experiments in this table are carried out by fine-tuning multilingual BERT model. Firstly, training on Basque projected data clearly outperforms direct model transfer strategies. Neither using English nor Spanish training data in the multilingual model is able to transfer the knowledge gained to Basque. Secondly, combining training data in different languages is beneficial for the performance on the target language. Joining English and Spanish training data with the Basque projected training set outperforms the Basque monolingual system consistently according to micro and macro F1. The best result on the Basque test set is achieved when training is done over Basque, English and Spanish data.

6.2. Slot filling

Flair related experiments were conducted using the Flair text classifier with the same hyperparameter settings as in intent classification experiments.
BERT fine-tuning was done with a learning rate of 2e-5 and a batch size of 16. The number of epochs was again optimized on the dev set, to the point were standard deviation of micro F1-score between the five runs was lower than 1. All systems presented in this section achieved the best performance after 5 epochs.

	EN	ES	EU
BERT monolingual	96.24	-	-
mBERT fine-tuning	96.37	88.84	76.77
Flair Embeddings	96.45	90.59	70.89
Previous results			
(Schuster et al., 2019)	94.81	82.96[6]	-

Table 5: Micro F1 results for Slot Filling.

Table 5 shows the results for the monolingual slot filling task. For English, all models achieve similar results, around 96% of accuracy. Instead, for Spanish dataset, Flair monolingual model slightly outperform the fine-tuned BERT multilingual model. In any case, note that both models outperform previous results achieved by (Schuster et al., 2019). On the contrary, Basque sequence labeler achieves best result with multilingual BERT. It is worth mentioning that even though the presence of the Basque in the Bert

[4] https://github.com/kyzhouhzau/BERT-NER/blob/master/conlleval.pl

[5] max-epochs 50, learning rate 0.1, minibatch size 64, and patience 3.

model is small, it is the model that performs best.

	micro F1	Macro F1
Monolingual		
EU (Projected)	76.77	57.41
Direct model transfer (cross-lingual)		
EN_{train}/EU_{test}	22.62	25.05
ES_{train}/EU_{test}	22.84	12.65
Multilingual training data		
$ES + EU_{train}/EU_{test}$	71.06	62.69
$EN + EU_{train}/EU_{test}$	74.04	65.77

Table 6: Slot filling results for direct transfer model experiments.

Table 6 shows the results achieved by the different training strategies for the slot filling task. All experiments in this table are carried out by fine-tuning multilingual BERT model. As in the previous task, we can conclude that training on Basque projected data clearly outperforms direct model transfer strategies. Neither using English nor Spanish training data is able to transfer the knowledge gained to Basque. This is not surprising considering the differences between the syntax of Basque and those of Spanish and English.

Finally, combining training data in different languages is partially beneficial for the sequence labeler, since the improvement is only achieved according to Macro F1. The main reason behind the lower Micro F1 is that monolingual data performs better over majority classes (*datetime*, *reminder/todo*, *reminder/noun*), but achieves poor results for classes with fewer examples (*Reminder/recurring-period*, *Reminder/reference*). The performance over these minority classes is better with the multilingual training, consequently the Macro F1 is higher for this approach. Basque monolingual system achieves the best accuracy or micro F1, that is, when training is done only over the automatically translated and projected Basque data.

7. Related work

Intent detection and **slot filling** are the main tasks involved in a task-oriented dialog process. Intent detection has been treated as a sentence classification problem by means of classifiers like Support Vector Machines (Chelba et al., 2003) or more recently neural network-based ones (Sarikaya et al., 2011). First approaches to deal with slot filling were HMM/CFG composite models (Wang et al., 2005) and Conditional Random Fields (CRF) (Raymond and Riccardi, 2007; Wang et al., 2011). Later, neural network-based approaches (Xu and Sarikaya, 2013; Mesnil et al., 2013) have been proposed to train both intent detection and slot filling together. (Xu and Sarikaya, 2013) propose a joint model for intent detection and slot filling based on Convolutional Neural Networks. (Mesnil et al., 2013) implemented and compared several Recurrent Neural Network architectures outperforming a CRF baseline substantially. (Liu and Lane, 2016) propose an attention-based neural network model for joint intent detection and slot filling. (Goo et al., 2018) consider that slot and intent are strongly related, and propose a slot gate that focuses on learning the relationship between intent and slot attention vectors.

Our work is closest to (Schuster et al., 2019). They deal with the problem of **developing task oriented dialog systems for less resourced languages**. They evaluate three different cross-lingual transfer methods to high-resource language to train models in low-resource languages: a) translating the training data. b) using cross-lingual pre-trained embeddings, and c) a novel method of using a multilingual machine translation encoder. They find that given several hundred training examples in the the target language, the latter two methods outperform translating the training data. The slot annotations are projected via the attention weights (Yarowsky et al., 2001). Our work differs from (Schuster et al., 2019) on various aspects. Firstly, our projection method is based on MT as well, but the alignment is performed by using the IBM Model 2 variant proposed by (Dyer et al., 2013). (Schuster et al., 2019) uses the the attention weights of the NMT for the alignment, but several authors (Koehn and Knowles, 2017; Ghader and Monz, 2017) have pointed out that attention-based alignments differ from human alignments. Secondly, we use different language models (Flair and BERT), and lastly, Basque is not addressed in their work.

Regarding **annotation projection** from source data to the (unlabeled) target data, there is work for several sequence labeling tasks such as POS tagging (Yarowsky et al., 2001; Duong et al., 2014), NER (Ehrmann et al., 2011; Kim et al., 2012; Ni and Florian, 2016), and parsing (McDonald et al., 2011). Most of the approaches are based on parallel corpora and only a few of them use MT systems (Tiedemann et al., 2014; Mayhew et al., 2017). Also, usually annotations are projected once translation process is finished (Yarowsky et al., 2001; Ni et al., 2017) by using unsupervised alignment models from statistical MT literature, such as IBM Models 1-6 (Brown et al., 1993; Och and Ney, 2003).

8. Conclusions

This work presents our research on the field of task-oriented dialog systems, focusing on intent classification and slot filling tasks. We have contributed to the field by researching methods that are applicable to less-resourced languages, with Basque as a case study. Specifically, we have compared a training data projection approach with direct model transfer strategies. Moreover, this work is to the best of our knowledge the first effort towards developing a task-oriented dialog system for Basque. We have created datasets for Basque and made them publicly available.

Our experiments show that training exclusively on Basque projected data achieves competitive results for both intent classification and slot filling tasks. Direct model transfer on the other hand, offers very low performance in either of the tasks. Hence, we conclude that the multilingual language model is not able to transfer the knowledge gained in a language (English or Spanish) to Basque in our datasets.

However, the multilingual language model benefits from multilingual training data. Best results in our experiments are achieved when rich-resource languages' data is added to Basque projected data for training.

There is still room for improvement. Even if the results obtained for Basque are competitive in terms of accuracy (81% for intent classification and 77% for slot filling) they

are still far from those achieved for English or Spanish. Our future research directions include improving the quality of our annotation alignment by splitting Basque inflections, on the one hand, and obtaining larger annotated datasets, by projecting English data to Spanish by means of NMT. Lastly, we have already talked about the limited presence of Basque in mBERT. A third line of work before us is to generate pre-trained multilingual language models containing a better balance of Basque data.

9. Acknowledgements

This work has been partially funded by the Basque Government Elkartek program (DL4NLP project, grant no. KK-2019/00045).

10. Bibliographical References

Adhikari, A., Ram, A., Tang, R., and Lin, J. (2019). Rethinking complex neural network architectures for document classification. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4046–4051.

Akbik, A., Blythe, D., and Vollgraf, R. (2018). Contextual string embeddings for sequence labeling. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 1638–1649.

Brown, P. F., Pietra, V. J. D., Pietra, S. A. D., and Mercer, R. L. (1993). The mathematics of statistical machine translation: Parameter estimation. *Computational linguistics*, 19(2):263–311.

Chelba, C., Mahajan, M., and Acero, A. (2003). Speech utterance classification. In *2003 IEEE International Conference on Acoustics, Speech, and Signal Processing, 2003. Proceedings.(ICASSP'03).*, volume 1, pages I–I. IEEE.

Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2019). BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Duong, L., Cohn, T., Verspoor, K., Bird, S., and Cook, P. (2014). What can we get from 1000 tokens? a case study of multilingual pos tagging for resource-poor languages. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 886–897.

Dyer, C., Chahuneau, V., and Smith, N. A. (2013). A simple, fast, and effective reparameterization of ibm model 2. In *Proceedings of NAACL-HLT*, pages 644–648.

Ehrmann, M., Turchi, M., and Steinberger, R. (2011). Building a multilingual named entity-annotated corpus using annotation projection. In *Proceedings of the International Conference Recent Advances in Natural Language Processing 2011*, pages 118–124.

Ghader, H. and Monz, C. (2017). What does attention in neural machine translation pay attention to? In *IJCNLP*.

Goo, C.-W., Gao, G., Hsu, Y.-K., Huo, C.-L., Chen, T.-C., Hsu, K.-W., and Chen, Y.-N. (2018). Slot-gated modeling for joint slot filling and intent prediction. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 753–757.

Hashemi, H. B., Asiaee, A., and Kraft, R. (2016). Query intent detection using convolutional neural networks. In *International Conference on Web Search and Data Mining, Workshop on Query Understanding*.

Henderson, M., Thomson, B., and Williams, J. D. (2014). The second dialog state tracking challenge. In *Proceedings of the 15th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*, pages 263–272, Philadelphia, PA, U.S.A., June. Association for Computational Linguistics.

Huang, Z., Xu, W., and Yu, K. (2015). Bidirectional lstm-crf models for sequence tagging. *arXiv preprint arXiv:1508.01991*.

Kim, S., Toutanova, K., and Yu, H. (2012). Multilingual named entity recognition using parallel data and metadata from wikipedia. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*, pages 694–702. Association for Computational Linguistics.

Koehn, P. and Knowles, R. (2017). Six challenges for neural machine translation. In *Proceedings of the First Workshop on Neural Machine Translation*, pages 28–39, Vancouver, August. Association for Computational Linguistics.

Liu, B. and Lane, I. (2016). Attention-based recurrent neural network models for joint intent detection and slot filling. *arXiv preprint arXiv:1609.01454*.

Ma, X. and Hovy, E. (2016). End-to-end sequence labeling via bi-directional lstm-cnns-crf. *arXiv preprint arXiv:1603.01354*.

Mayhew, S., Tsai, C.-T., and Roth, D. (2017). Cheap translation for cross-lingual named entity recognition. In *Proceedings of the 2017 conference on empirical methods in natural language processing*, pages 2536–2545.

McDonald, R., Petrov, S., and Hall, K. (2011). Multi-source transfer of delexicalized dependency parsers. In *Proceedings of the conference on empirical methods in natural language processing*, pages 62–72. Association for Computational Linguistics.

Mesnil, G., He, X., Deng, L., and Bengio, Y. (2013). Investigation of recurrent-neural-network architectures and learning methods for spoken language understanding. In *Interspeech*, pages 3771–3775.

Ni, J. and Florian, R. (2016). Improving multilingual named entity recognition with Wikipedia entity type mapping. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1275–1284, Austin, Texas, November. Association for Computational Linguistics.

Ni, J., Dinu, G., and Florian, R. (2017). Weakly supervised cross-lingual named entity recognition via effective annotation and representation projection. *arXiv preprint*

arXiv:1707.02483.

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

Raymond, C. and Riccardi, G. (2007). Generative and discriminative algorithms for spoken language understanding. In *Eighth Annual Conference of the International Speech Communication Association*.

Sarikaya, R., Hinton, G. E., and Ramabhadran, B. (2011). Deep belief nets for natural language call-routing. In *2011 IEEE International conference on acoustics, speech and signal processing (ICASSP)*, pages 5680–5683. IEEE.

Schuster, S., Gupta, S., Shah, R., and Lewis, M. (2019). Cross-lingual transfer learning for multilingual task oriented dialog. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 3795–3805, June.

Tiedemann, J., Agić, Ž., and Nivre, J. (2014). Treebank translation for cross-lingual parser induction. In *Proceedings of the Eighteenth Conference on Computational Natural Language Learning*, pages 130–140.

Tur, G., Hakkani-Tür, D., and Heck, L. (2010). What is left to be understood in atis? In *2010 IEEE Spoken Language Technology Workshop*, pages 19–24. IEEE.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, Ł., and Polosukhin, I. (2017). Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Wang, Y.-Y., Deng, L., and Acero, A. (2005). Spoken language understanding. *IEEE Signal Processing Magazine*, 22(5):16–31.

Wang, Y., Deng, L., and Acero, A. (2011). Semantic frame-based spoken language understanding. *Spoken Language Understanding: Systems for Extracting Semantic Information from Speech*, pages 41–91.

Xu, P. and Sarikaya, R. (2013). Convolutional neural network based triangular crf for joint intent detection and slot filling. In *2013 IEEE Workshop on Automatic Speech Recognition and Understanding*, pages 78–83. IEEE.

Yarowsky, D., Ngai, G., and Wicentowski, R. (2001). Inducing multilingual text analysis tools via robust projection across aligned corpora. In *Proceedings of the first international conference on Human language technology research*, pages 1–8. Association for Computational Linguistics.