# Cross-lingual intent classification in a low resource industrial setting

**Talaat Khalil*** **Kornel Kiełczewski*** **Georgios Christos Chouliaras**
**Amina Keldibek** **Maarten Versteegh**
Booking.com Amsterdam
{talaat.khalil,kornel.kielczewski}@booking.com

## Abstract

This paper explores different approaches to multilingual intent classification in a low resource setting. Recent advances in multilingual text representations promise cross-lingual transfer for classifiers. We investigate the potential for this transfer in an applied industrial setting and compare to multilingual classification using machine translated text. Our results show that while the recently developed methods show promise, practical application calls for a combination of techniques for useful results.

## 1 Introduction

Classifying the intent represented in a customer message is a core functionality of natural language understanding applications in customer service. Once the intent of the customer's message is understood and various entities are resolved, actions can be triggered, such as automatically replying to the customer, routing the message to the right customer service representative, or asking the customer for more information.

The rise of automation in customer service and the growing customer expectation of immediate answers is generating increasing demand for accurate intent detection across many languages. Recent advances in natural language processing have enabled impressive performance on text classification tasks, provided that large-scale labelled data are available. In industrial settings, this is often not the case, or at least not for all languages for which intent detection is required.

This paper studies what we believe to be a common case in industry, in which a small amount of labeled data are available for a single language, say English, but neither data nor labels are available for other languages. This setting poses an interesting constraint on the application of modern machine learning techniques, in that much fewer in-domain data are available for the task than is generally required. We believe that the findings in this paper can be generalized outside our specific industrial domain and apply in general to the challenge of enabling classification in a target language based on supervision in a source language.

This paper investigates the application of two techniques from the literature to this problem. The first technique is an out-of-domain machine translation system, as can be assumed to be generally available, applied to the source language data set in order to generate supervised data in the target language. The second technique that we investigate is transfer learning based on multi-lingual document representations. We study the results of applying these techniques, as well as the combination of both, in a number of experiments.

## 2 Related work

This study is broadly positioned in the application of cross-lingual transfer, with special focus on low resource applications. We study two approaches to this problem: machine translation and transfer through multilingual text representations.

One approach to NLP tasks in multilingual settings where the target language has scarce resources relies on machine translation. There are two popular methods: translating target language to English and vice versa, translating English to target language (Garca et al., 2012; He et al., 2013). While straightforward, a disadvantage of this approach is that building a reliable machine translation system with few resources in the target language is challenging (Upadhyay et al., 2018).

Enriched word embeddings are used to solve

* These authors contributed equally.



*Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing*, pages 6419–6424,
Hong Kong, China, November 3–7, 2019. 

various (monolingual) NLP tasks including intent classification (Kim et al., 2016). Multilingual representations of characters, words or documents can be used to solve multilingual NLP tasks. (Klementiev et al., 2012) present a representation for single words for a pair of languages that preserves semantic similarity. This approach relies on parallel data sets, a challenge to their scalability. To address this issue (Søgaard et al., 2015) propose an inter-lingual representation of words obtained with inverted indexing from common subsets of Wikipedia articles on a larger set of languages.

A recent approach is the construction of alignments between monolingual word representations for different languages into a single embedding space (Ammar et al., 2016; Alaux et al., 2018; Conneau et al., 2017; Schuster et al., 2019).

Recent studies evaluated shared sentence representations from multiple languages using both unsupervised learning on monolingual corpora (Lample and Conneau, 2019; Eriguchi et al., 2018) and supervised learning using parallel data (Artetxe and Schwenk, 2018). Advances in generative pretraining models like the Transformer (Vaswani et al., 2017), GPT (Radford et al., 2018), and BERT (Devlin et al., 2018) models make this approach even more promising. In (Lample and Conneau, 2019) authors successfully apply these techniques and propose methods for cross-lingual language modelling, unsupervised and supervised, outperforming previous best approaches.

In this paper, we study the capability of transfer from a source language to a target language through machine translation and pretrained multilingual document representations.

## 3 Data sets

The data set for the source language ("EN") consists of 48,875 English chat messages from customers to the customer service department of a large e-commerce website. The messages typically express intents from customers to get information or make changes on an existing transaction, e.g. cancelling a past transaction, modifying it in a certain way, or enquiring about information related to the transaction or the product. A test set for the target language ("FR") of 4,336 messages was similarly collected from the production system of the conversational agent. The median number of words in the source language messages is 16, the 99$^{th}$ percentile is at 93 words.

The messages were anonymized by replacing email addresses, transaction id's, credit card and telephone numbers and other personally identifiable information with token placeholders. After this process, The messages were manually labeled on 17 intents, covering the majority of incoming requests. Each message received one or more intent labels. The distribution of intents is shown in Figure 1 showing that the English and French sets have roughly the same label distribution, although there are some differences.

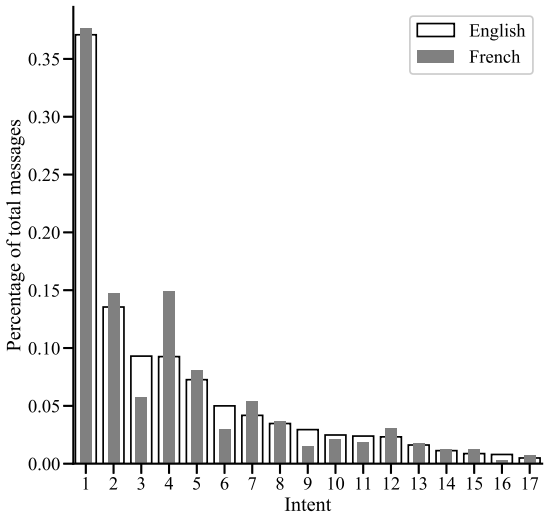


Figure 1: Intent distribution for English and French.

To support the investigation of using machine translation, an additional data set ("MT") was constructed by translating the EN data set using an in-house neural machine translation model (Levin et al., 2017). The machine translation model was trained on the open parallel corpus (Tiedemann, 2012), product descriptions, and fine-tuned for user reviews. While the product descriptions apply to the same industry as the chat messages we are investigating here, the training set for the machine translation model contains no documents from customer service interactions, so we consider it to be an out-of-domain system for the purpose of this study, as might be available to any investigator.

The EN data were split into fixed train, development, and test sets. The same splits were maintained for the translated (MT) data set. The FR data were used only for testing. Table 1 shows the volume of each of the data sets.

| Language | Train | Dev | Test |
|---|---|---|---|
| EN | 39,094 | 4,887 | 4,894 |
| MT | 39,094 | 4,887 | - |
| FR | - | - | 4,336 |

Table 1: Data set sizes for each approach.

## 4 Experiments

We study the effects of two approaches to cross-lingual transfer for intent classification; machine translation and transfer from multilingual document representations. We conduct three main experiments; one each for the two approaches and one that combines both approaches.

### 4.1 Machine translation

Machine translation for cross-lingual transfer is assessed by comparing a classifier trained and tested on the source language (EN) with one trained on the machine translated samples (MT) and tested on the target language (FR).

We evaluate two text classification architectures. The first architecture is a linear support vector machine (SVM) with bag-of-ngram input features and $\ell_1$- and $\ell_2$-regularization, a classic text classification setup (Sebastiani, 2002). A combination of unigram, bigram, and trigram features with tf-idf weighting is used.

The second text classification architecture is a convolutional neural network (CNN) similar to (Kim, 2014). The system represents input documents as sequences of word embeddings (Mikolov et al., 2018; Grave et al., 2018) that are passed through a single convolutional layer followed by a densely connected layer. The network was trained using the Adam (Kingma and Ba, 2015) optimizer and the hyperparameters of the network (size and amount of kernels, size of pre-final layer, dropout- and $\ell_1$-, $\ell_2$-regularization strength) were found using Bayesian optimization (Snoek et al., 2012). These text classification systems were chosen because both are commonly used in industry.

To assess machine translation for cross-lingual transfer, we compare two settings for each of the classifiers. In the first setting, both training and test data come from the source language (EN). In the second setting, training data is machine translated (MT) and test data is drawn from the target language (FR). If machine translation is a feasible strategy for transfer, then the results for the second setting should be close to those for the first.

### 4.2 Multilingual document embeddings

To investigate the use of multilingual document embeddings for cross-lingual transfer, we build a set of classifiers based on representations from the base multilingual BERT model (Devlin et al., 2018). We use the aggregate document representation of the input message and add a single output layer with sigmoid activations for multi-label classification, as recommended in (Devlin et al., 2018). With this system we study two setups.

In the first setup a classifier is trained on the source language (EN) only, and evaluated on both the source (EN) and the target language (FR). The hypothesis here is that if the multilingual representations from BERT allow for lossless transfer between source and target language, then the performance on the target language will be close to the performance on the source language.

In the second setup, we combine the approaches of machine translation and multilingual document representation. We train a similar classifier as above with input representations from BERT on both the source language (EN) data and the machine translated target language (MT) data. A separate classifier is trained only on the machine translated (MT) training set. We compare the performance of these two classifiers on the target (FR) test set. If the architecture facilitates task-specific transfer learning then there will be a difference in the performance between the two classifiers.

Lastly we compare the classifier that uses multilingual document representations and machine translated data with the classifiers that use only either of those approaches. If there is a cumulative effect from the combination of these approaches, then the first will outperform both of the latter.

## 5 Results

For the application described in this paper, false positive identifications of a class are worse than false negatives. Therefore the classifiers are calibrated on the development set to a precision close to 0.90 and maximum recall. This is a common practice in applications in which there is an asymmetry in the cost of errors. Table 2 shows the main results. Note that not all systems could achieve the required precision.

For assessing the feasibility of machine translation for cross-lingual transfer, we compare the re-

| Test | Train | System | P | R | F1 |
|---|---|---|---|---|---|
| EN | EN | svm | .91 | .60 | .72 |
| | EN | cnn | .90 | .67 | .77 |
| | EN | bert | .91 | **.71** | .80 |
| | EN+MT | bert | .91 | **.71** | .80 |
| FR | MT | svm | .86 | .43 | .57 |
| | MT | cnn | .86 | .47 | .61 |
| | EN | bert | .74 | .51 | .61 |
| | MT | bert | .86 | .62 | .72 |
| | EN+MT | bert | .86 | **.63** | .73 |

Table 2: Test set results for the experiments.

sults of the linear SVM and the CNN. For both of these classifiers, the source (EN) to source (EN) setting outperforms the translated (MT) to target (FR) setting, by 0.17 to 0.2 recall and 0.15 to 0.16 F1-score. The latter results are, however, substantially better than chance for both classifiers. It is interesting to note that the results of the SVM do not lag far behind those of the CNN, both for EN and FR, indicating that the machine translation approach is feasible even for the SVM.

Next, the capability of multi-lingual document representations to provide cross-lingual transfer is tested by comparing the performance on EN and FR for the BERT model trained solely on EN. For the FR test case, this model could not be calibrated for a precision $\geq 0.9$, so we can only compare the F1-scores. Here again, the EN setting shows better performance than the FR setting (F1 of 0.80 versus 0.61), indicating that the transfer incurs a loss in performance, but is not entirely without meaning.

Lastly, the BERT system trained on translated (MT) data shows the combination of the approaches. The evaluation on target data (FR) shows that this combination achieves higher performance than all of the above settings, outperforming MT only (0.62 recall versus 0.47 and 0.43), as well as only transfer from multilingual embeddings (0.72 F1 versus 0.61). Adding EN data to this combination model improves performance slightly (0.63 versus 0.62 recall), but these extra data do not seem to have a large effect on the performance of the system. In addition, this latter BERT model, trained on both EN and MT, when tested on the EN test set performs on par with the BERT model trained only on EN data. The implication of this is that the MT data do not introduce a degradation of the model. It is interesting to note that the results of this setting on the target language are higher than those of the SVM on the source language.

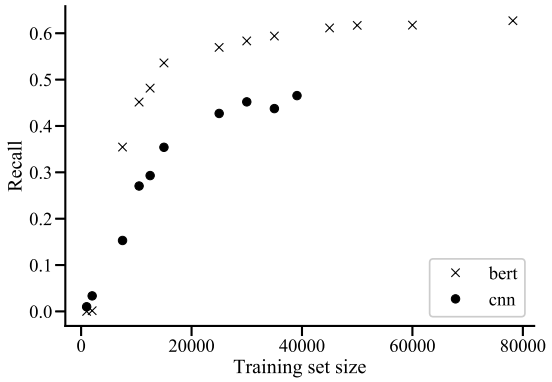


Figure 2: Recall as a function of the size of the training data. This figure shows the recall of the BERT model trained on both EN and MT and the recall of the CNN model trained on MT.

To evaluate the dependence of the BERT system with source and translated data on the amount of data available for fine-tuning the BERT model, we look at the performance as a function of the data set size and compare it in this respect to the CNN. Figure 2 shows that both the BERT model and the CNN performances converge on about 10,000 to 25,000, indicating that only a relatively small amount of labeled data is necessary.

## 6 Discussion

In this paper we studied two techniques for cross-lingual transfer in a typical industrial setting, with small amounts of labeled data. The results show that both machine translated data and multilingual document representations are decent strategies for cross-lingual transfer for intent detection on short chat messages. Both approaches incur a significant loss when compared to the source-to-source classification results, showing that there is still room for improvement before "zero-shot" transfer from multilingual document representations is viable as a stand-alone approach to cross-lingual transfer.

The combination of both techniques performs competitively and the observation that this setting outperforms the SVM in the source-to-source setting illustrates the advances that have been made in recent years in both machine translation and document representation.

It is an open question on whether these results generalize to other language pairs. The languages studied here are both western Indo-European languages so the results obtained here may not apply to pairs of distant languages. In addition, the label distributions between the source and target language in our experiments were fairly closely aligned, which may not apply to other use cases. In general, more investigation is needed to assess how well the current generation of multi-lingual document representations supports cross-lingual transfer and if there are differences in how well the representations between languages align.

It is promising to note though that the results in this study can be obtained with data set sizes that are generally available in industrial settings.

## Acknowledgements

The authors wish to thank Elena Sokolova for useful discussions, Pavel Levin for his comments on an early version of this paper and Nam Pham for his help with the hyperparameter optimization.

## References

Jean Alaux, Edouard Grave, Marco Cuturi, and Armand Joulin. 2018. Unsupervised hyperalignment for multilingual word embeddings. *CoRR*, abs/1811.01124.

Waleed Ammar, George Mulcaire, Yulia Tsvetkov, Guillaume Lample, Chris Dyer, and Noah A. Smith. 2016. Massively multilingual word embeddings. *CoRR*, abs/1602.01925.

Mikel Artetxe and Holger Schwenk. 2018. Margin-based parallel corpus mining with multilingual sentence embeddings. *CoRR*, abs/1811.01136.

Alexis Conneau, Guillaume Lample, Marc'Aurelio Ranzato, Ludovic Denoyer, and Hervé Jégou. 2017. Word translation without parallel data. *CoRR*, abs/1710.04087.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Akiko Eriguchi, Melvin Johnson, Orhan Firat, Hideto Kazawa, and Wolfgang Macherey. 2018. Zero-shot cross-lingual classification using multilingual neural machine translation. *CoRR*, abs/1809.04686.

F. Garca, L. F. Hurtado, E. Segarra, E. Sanchis, and G. Riccardi. 2012. Combining multiple translation systems for spoken language understanding portability. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, pages 194–198.

Edouard Grave, Piotr Bojanowski, Prakhar Gupta, Armand Joulin, and Tomas Mikolov. 2018. Learning word vectors for 157 languages. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

X. He, L. Deng, D. Hakkani-Tur, and G. Tur. 2013. Multi-style adaptive training for robust cross-lingual spoken language understanding. In *2013 IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 8342–8346.

J. Kim, G. Tur, A. Celikyilmaz, B. Cao, and Y. Wang. 2016. Intent detection using semantically enriched word embeddings. In *2016 IEEE Spoken Language Technology Workshop (SLT)*, pages 414–419.

Yoon Kim. 2014. Convolutional neural networks for sentence classification. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*.

Diederik P. Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. *CoRR*, abs/1412.6980.

Alexandre Klementiev, Ivan Titov, and Binod Bhattarai. 2012. Inducing crosslingual distributed representations of words. In *Proceedings of COLING 2012*, pages 1459–1474, Mumbai, India. The COLING 2012 Organizing Committee.

Guillaume Lample and Alexis Conneau. 2019. Cross-lingual language model pretraining. *CoRR*, abs/1901.07291.

Pavel Levin, Dhanuka. Nishikant, Talaat Khalil, Fedor Kovalev, and Maxim Khalilov. 2017. Toward a full-scale neural machine translation in production: the Booking.com use case. In *Proceedings of Machine Translation Summit XVI*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training.

Tal Schuster, Ori Ram, Regina Barzilay, and Amir Globerson. 2019. Cross-lingual alignment of contextual word embeddings, with applications to zero-shot dependency parsing. *CoRR*, abs/1902.09492.

Fabrizio Sebastiani. 2002. Machine learning in automated text categorization. *ACM computing surveys (CSUR)*, 34(1):1–47.

Jasper Snoek, Hugo Larochelle, and Ryan P. Adams. 2012. Practical Bayesian optimization of machine

learning algorithms. In *Proceedings of the 25th International Conference on Neural Information Processing Systems - Volume 2*, NIPS'12, pages 2951–2959, USA. Curran Associates Inc.

Anders Søgaard, Željko Agić, Héctor Martínez Alonso, Barbara Plank, Bernd Bohnet, and Anders Johannsen. 2015. Inverted indexing for cross-lingual NLP. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1713–1722, Beijing, China. Association for Computational Linguistics.

Jörg Tiedemann. 2012. Parallel data, tools and interfaces in opus. In *Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12)*, Istanbul, Turkey. European Language Resources Association (ELRA).

Shyam Upadhyay, Manaal Faruqui, Gokhan Tur, Dilek Hakkani-Tur, and Larry Heck. 2018. (almost) zero-shot cross-lingual spoken language understanding. In *Proceedings of the IEEE ICASSP*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *CoRR*, abs/1706.03762.