Bootstrapping Multilingual Intent Models via Machine Translation for Dialog Automation

Nicholas Ruiz, Srinivas Bangalore, John Chen
Interactions, LLC
Murray Hill, NJ, USA
{nruiz,sbangalore,jchen}@interactions.com

Abstract

With the resurgence of chat-based dialog systems in consumer and enterprise applications, there has been much success in developing data-driven and rule-based natural language models to understand human intent. Since these models require large amounts of data and in-domain knowledge, expanding an equivalent service into new markets is disrupted by language barriers that inhibit dialog automation.

This paper presents a user study to evaluate the utility of out-of-the-box machine translation technology to (1) rapidly bootstrap multilingual spoken dialog systems and (2) enable existing human analysts to understand foreign language utterances. We additionally evaluate the utility of machine translation in human assisted environments, where a portion of the traffic is processed by analysts. In English→Spanish experiments, we observe a high potential for dialog automation, as well as the potential for human analysts to process foreign language utterances with high accuracy.

1 Introduction

With the present advances in natural language understanding and speech recognition technologies, unprecedented opportunities have been created for realizing natural and sophisticated human-machine conversations to accomplish routine tasks. There has been a resurgence of speech/text-based conversation systems spanning multiple platforms, such as interactive voice recordings, chat, and SMS, owing to the availability of communication platforms that make it convenient to configure human-machine conversations. The potential opportunities of speech/text driven human-machine systems, or *virtual agents*, can only be realized if the user's requests are *understood* by the virtual agent and acted upon appropriately.

For practical applications, such conversational agents and speech/text analytics systems, the meaning of a sentence may be approximated as one or more actionable labels, or *intents*, associated with the input utterance. In such cases, the natural language understanding (NLU) task is modeled as an *intent classification* problem. Although ambiguity is present in natural language, data-driven NLU systems have been successful in modeling user intents in many application domains.

Many commercial and enterprise applications service customers from different geographic locations and varying language proficiencies, requiring multilingual NLU for human-machine interaction. In order to deploy an intent classification system for a new language a new set of labeled training data is conventionally required. This data is often unavailable before a solution is deployed, instead requiring a human-driven dialog system depending on intent analysts or live agents. In time, a sufficient amount of production data may be collected to build a data-driven intent model; however this approach is expensive to operate and ignores valuable knowledge present in other languages that could be used to build an initial model.

In this paper, we evaluate the use of machine translation (MT) as a tool to bridge the knowledge present in one or more intent models for the creation of an intent classifier in a target language. Given MT's capability to translate the content of

Pérez-Ortiz, Sánchez-Martínez, Esplà-Gomis, Popović, Rico, Martins, Van den Bogaert, Forcada (eds.)
Proceedings of the 21st Annual Conference of the European Association for Machine Translation, p. 303–308
Alacant, Spain, May 2018.

utterances in a source language to a target language, our goal is to minimize the number of language proficient intent analysts needed to support a production-scale multilingual dialog system in the absence of target language data.

The remainder of the paper is organized as follows: in Section 2 we discuss the details of the intent classification system and present the NLU model, including our two MT architectures in a multilingual spoken dialog system. In Section 3 we outline our experiment and describe our ASR, MT, and NLU models. In Sections 4-7, we evaluate the ASR, MT, and NLU model performance. In Section 8 we evaluate human agents' ability to label the intents of translated user utterances and summarize our findings in Section 10.

2 NLU for Customer Care

A single utterance is tagged with three types of labels: *intents*, *entities*, and *conversational handlers*. Intents are domain-specific labels such as SALES, TECH ASSISTANCE, and BILLING. *Entity* labels represent the names of products or services mentioned by the user. These include specific models of smart phones or subscription services. Conversational handlers are labels which are similar to speech acts to guide the conversation. For example, LIVE AGENT, CONFUSED, or FOREIGN LANGUAGE. In our experiments, intents and entities are labeled as "session variables" (SV), while conversation handles are partitioned into "task names" (TN) and "event names" (EN). Examples of each in our experiments are shown in Table 1.

A joint SVM classifier is trained by concatenating the TN, EN, and SV labels into a unique label. The model comprises a set of binary SVM classifiers, with each classifier predicting if the input is assigned or not assigned to a particular label type. For a given input utterance x, the joint label is computed as:

$$y^* = \underset{y \in \langle tn, sv, en \rangle}{\arg\max} \; F_y(x, y). \tag{1}$$

The feature set *F* may comprise *application context*, *conversation context*, and the *utterance*. We use n-gram word-level features in this experiment.

2.1 Confidence Measures

We boost the accuracy of our intent models by using human analysts' predictions on unconfident decisions made by the classifier. We obtain prediction probabilities from the classifier by computing the sigmoid on the scores output for each label by the SVM classifier, computed as:

$$P(y^*) = \frac{1}{1 + \exp(F_y(x, y^*))}. \tag{2}$$

The confidence measure is obtained by computing the ratio of the probabilities of the first and second best labels assigned by the classifier:

$$cf(y^*) = \frac{P(y^*)}{P(y^{*-1})}. \tag{3}$$

The rejection threshold is empirically determined to maximize the accuracy of the human-assisted solution while minimizing human labeling costs.

Task Names (TN)	Session Variables (SV)	Event Names (EN)
COMPLAINT	ACCOUNT action	ANGRY
ENGLISH	ACTIVATE product	DON'T KNOW
FOREIGN	ADD service	GARBLED
NONE	APPOINTMENT type	LIVE AGENT
QUESTION	BILLING AUTO PAY	NOISE
...	BILLING DETAILS	NO MATCH
	PAY BILL service	NONE
	CHANGE ADDRESS	THANK YOU
	DISCONNECT service	...
	...	
10	707	18

Table 1: Examples and counts of Spanish intent labels by category and language for the "How may I help you?" dialog state. Concatenating the labels yields 833 distinct intent annotations.

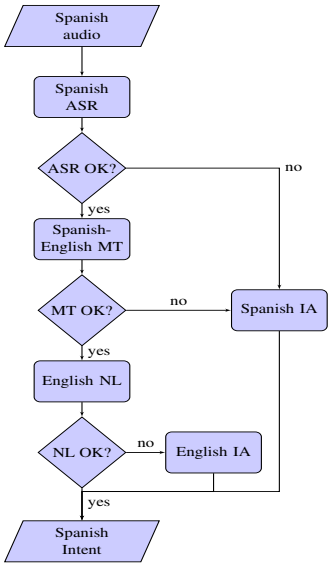

Figure 1: Online Spanish→English bootstrap architecture. Spanish audio is translated into English and processed by an English intent classifier. If ASR or MT scores are not confident, a Spanish intent analyst (IA) labels the segment. If the English intent model is not confident, an English IA labels it.

2.2 Multilingual Bridging via Machine Translation

In the context of dialog systems, MT can be used in one of two ways: (1) translating real-time target

Data set	# utts	# words	# unique labels
English training	6.5M	40.6M	623
Spanish training	0.8M	4.6M	623
Spanish test (ASR)	1007	4696	178
Spanish test (human)	1007	5153	178

Table 2: Utterance, word, and distinct label counts for the training and test data used in our experiments.

language data into the source language and predicting the intent with a source language intent classification model (cf. Fig. 1); or (2) translating source language data offline into the target language and training a target language intent classification model. In the first scenario, utterances with high translation quality may be processed by an analyst that only speaks the source language, if the NLU confidence score is too low.

3 Experimental setup

We evaluate the efficacy of bootstrapping a Spanish intent classifier using the data and underlying models from an English spoken language dialog system. The data set consists of customer voice responses to the message "How may I help you?" in the customer's native language at the beginning of a phone call to an Interactive Voice Response (IVR) system. We assume that no Spanish intent data is available during training time and evaluate the performance of our bootstrapped Spanish models against an intent model trained with a standard training set. Table 2 lists the data used in our experiment, The training data consists of 5-best ASR hypotheses on audio segments for English and Spanish. We assume that the intent labels covered by the target Spanish model are the same as the English model. Although most of the intent labels overlap one another there is a subset of $\langle tn, sv, en \rangle$ intent triples that do not overlap. We discard the non-training examples with non-overlapping triples from each data set, reducing the number of unique labels in both data sets to 623 (cf. Table 1). As a result, 2.94% of the English training examples and 1.95% of the Spanish training examples are discarded, respectively. Model performance is evaluated on a test set of 1007 Spanish audio segments randomly extracted from production logs and labeled by Spanish-speaking intent analysts (IAs) and verified by a supervisor.

3.1 Automatic Speech Recognition

Our Spanish ASR system consists of an n-gram language model and hybrid DNN acoustic model trained with the cross-entropy criterion followed by the state-level Minimum Bayes Risk (sMBR) objective. We use sequence-training with smoothing and speed-perturb the training data. The acoustic model has general-phone and head-body-tail based digit-specific triphones. The training data consists of 500K training utterances (around 1000 hours of audio) without speech perturbations and about 40K unique vocabulary words.

3.2 Machine Translation

We use a conventional neural machine translation (NMT) sequence-to-sequence encoder-decoder with attention architecture (Bahdanau et al., 2015; Luong and Manning, 2015; Sennrich et al., 2016) commonly used by MT practitioners. The NMT models were trained with parallel English-Spanish data from Europarl v7, CommonCrawl, and WMT News Commentary v8 from the WMT 2013 evaluation campaign (Bojar et al., 2013), as well as the TED talks from IWSLT 2014 (Cettolo et al., 2014). The training data has a shared vocabulary size of 89,500 words after byte-pair encoding (Sennrich et al., 2016). The model is trained for 20 epochs with two bidirectional LSTM encoding and decoding layers with 512 units. In this experiment we assume to have no in-domain parallel data.

For the offline English→Spanish model, we translate the English intent model's training data (6.5 million utterances) into Spanish using our baseline NMT system. The number of words in the translated data set remains roughly the same. The translated outputs are used to train a bootstrapped Spanish intent classifier, using the same training parameters as the native English model. The ASR outputs from the test set are processed by the bootstrapped Spanish intent classifier. For the real-time Spanish→English model, insert punctuation and apply truecasing to the ASR outputs from the test set and translate the outputs with our Spanish→English baseline NMT system. We strip the punctuation and lowercase the machine translated output and pass it through the native English intent classifier.

3.3 Intent classification

Our intent classifiers are trained using an implementation of SVMs in SCIKIT-LEARN[1], using the approach described in Section 2. We evaluate the

[1] http://scikit-learn.org

Translation	BLEU ↑	TER ↓	Length
ASR outputs	42.5	51.2	94.1
ASR outputs (+PE)	48.0	44.6	96.3
Human transcripts	51.8	41.3	111.4

Table 3: Machine translation quality measured in BLEU, TER, and utterance length, evaluated against post-edited translations of human transcripts.

performance of an intent classification model by plotting an *error-rejection curve*, which measures the error rate of the intent classifier as the number of utterances that are processed by the intent analysts increases. For example, a 10% rejection rate corresponds means that only 90% of the test set is evaluated by the model.

We compare the results of each bootstrapped NLU approach with a native Spanish intent model trained on the held-out Spanish training data. We additionally repeat the experiment with the human transcripts to measure the difference in intent classification error that may be explained by ASR. Error-rejection curves for each intent model are shown in Fig. 2 and the scores at 0%, 10%, and 20% rejection are shown in Table 4.

4 ASR performance

We use SCLITE from the NIST Speech Recognition Scoring Toolkit[2] to compute the word error rate (WER) and utterance error rate. After further clean-up and adjudication, we observe a 33.2% WER, with 60.0% of the utterances containing errors (32% substitutions, 22% deletions, 28% insertions). A majority of the substitution errors were confusions between singular and plural (e.g. *problemas→problema*) and articles (e.g. *del→de*). Other errors included phonetic confusions (e.g. *cuenta→fuenta*), named entity misrecognitions (e.g. *HBO→yo*), and a high frequency of dropped articles (e.g. *de*, *la*, *a*) caused by speaker under-articulation. Of these types of errors, the most detrimental are substitutions of named entities, verbs, and nouns that are not lemmatization errors. Another issue driving up the WER score caused by the IVR system truncating audio longer than four seconds to reduce latency.

5 Machine Translation quality

In order to assess the translation quality, we post-edit the translations of the human-transcribed utterances and report the case-insensitive BLEU and TER scores in Table 3. ASR errors increase the required translation edits by 10%, from 41.3% edits to 51.2%. The primary sources of errors are incomplete sentences, lack of punctuation; lexical mistranslations of key words: (e.g. *equipo* (*equipment*)→*team*; *dirección* (*address*)→*direction*; *reclamo* (*complaint*)→ *claim*); ambiguous translations (e.g. *factura→bill/invoice*); and duplicated words during translation (e.g. *payment arrangements→payment payment*). Many of these issues are due to lack of in-domain MT training data and low tolerance of ASR errors. Highly repetitive errors indicate that an automatic post-editing system could substantially improve the translation system's quality. Table 3 also shows that post-edited translations of ASR outputs are substantially worse than those of the human transcripts (41.3% TER difference), showing that ASR errors are exacerbated through translation.

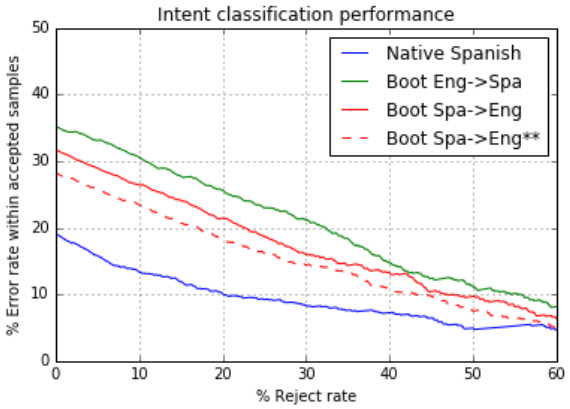

Figure 2: Error-rejection curves for bootstrapped intent classification performance on Spanish ASR outputs, versus a native Spanish model. Results are reported for intent classifier predictions on ASR outputs. Spa→Eng**: performance on post-edited MT outputs.

6 Native NLU performance

Our reference native Spanish intent classification model is trained on ASR outputs since an insufficient amount of human-transcribed intent modeling data is available. From Table 4, we see that at 0% rejection, the native model yields a 19.2% classification error, while the human intent analysts (IAs) yielded a 11.0% error while listening directly to the audio. At the same time, if the IAs are presented with the ASR outputs, they produce an error rate of 24.4%. This demonstrates the intent model's ability to tolerate a certain degree of ASR errors by being trained on ASR errors.

[2]https://www.nist.gov/itl/iad/mig/tools

Configuration	Reject 0%	10%	20%
Native Spanish (ASR)	19.2	13.5	10.3
Native Spanish (human)	19.1	12.3	7.4
Boot Eng→Spa (ASR)	35.3	30.7	25.8
Boot Eng→Spa (human)	31.6	26.3	21.4
Boot Spa→Eng* (ASR)	31.8	26.5	21.4
Boot Spa→Eng* (human)	33.4	26.7	20.2
Boot Spa→Eng* (ASR+PE)	28.3	23.6	18.3
Spanish IA (audio)	11.0	10.6	9.5
Spanish IA (human)	14.9	13.1	10.9
Spanish IA (ASR)	24.4	21.6	17.6
English IA (ASR+MT)	33.9	30.6	26.1
English IA (ASR+MT+PE)	25.9	22.8	18.3

Table 4: Intent classification performance by machine learning models and human intent analysts (IAs) at 0%, 10%, and 20% rejection.

Of the 1007 test utterances, 152 utterances have both ASR errors and were classified incorrectly, although they were labeled properly by IAs, comprising 15.1% of the 19.2% intent classification errors. Of the utterances with ASR errors that cause an intent classification error, six were cases where ASR failed to produce a hypothesis, 26 were cases of audio truncation, and 39 utterances containing only ASR substitution errors. The latter cases often caused underspecification during intent classification (e.g. LOWER MY BILL→BILLING PROBLEM; MAKE A PAYMENT→BILLING). In isolation, insertion and deletion ASR errors did not have a significant impact on NLU.

Evaluating NLU performance on human transcripts, we observe improvements that rival human labeling performance at 10% rejection and above. At 20% rejection, the model significantly outperforms human labeling.

7 Bootstrapped NLU performance

While the native Spanish model has a 13.5% error rate at 10% rejection when processing ASR hypotheses, the bootstrapped models have double the error rate due to the use of out-of-domain machine translation. On ASR hypotheses, the English→Spanish model yields an error rate of 30.7%, while the Spanish→English model yields 26.5% at 10% rejection.

To better understand how machine translation further corrupts the bootstrapped Spanish→English performance, we compare the errors it makes to the Native Spanish model. In Table 5, we group the NLU errors by whether they are present in the Native Spanish model, the bootstrapped model, or both. 14.2% of the test set are

Native	Spa→Eng	ASR	ASR+PE
+	+	66.6%	69.6%
+	-	14.2%	11.2%
-	+	1.5%	2.0%
-	-	17.7%	17.2%

Table 5: Comparison of Native Spanish intent model to bootstrapped Spanish→English models on ASR outputs. +/- indicate that whether the corresponding model's prediction was correct.

examples where the Native Spanish model makes a correct prediction, but the Spanish→English model yields errors. By comparing to post-edited ASR+MT data, only 3% of those errors are directly attributed to ASR errors. The 11% of Spanish→English model-specific errors are mostly attributed to intent underspecification. For example, 60% of the ACCOUNT errors are NULL misclassifications. For billing issues, two-thirds of the errors are semantically similar misclassifications, such as PAYMENT, LOWER MY BILL, and BILL DETAILS.

8 Analyst performance on translated text

Finally, we measure IA labeling performance on translated utterances. In our conventional scenario, intent analysts listen to audio segments in their native language and provide an intent label. Instead, we replace the original audio with machine translated or translation post-edits of ASR hypotheses. Fig. 3 provides error-rejection curves for IAs, with error rates at 0%, 10% and 20% rejection reported in Table 4.

We first assess the labeling loss when humans annotate ASR outputs in the absence of audio. Although their error rate increases from 11.0% to 24.4% when annotating ASR transcripts, their performance on human transcripts is within 5% of listening directly to the audio at 0% rejection. As the rejection rate increases, the difference becomes negligible. As we introduce Spanish→English machine translation, we observe that the labeling error increases from 24.4% to 33.9% on ASR, which is incidentally worse than the Spanish→English intent classification model's accuracy (31.8%)! However, the IA labeling error rate drops to 25.9% on post-edited MT outputs – only a 1.5% increase in NLU errors caused by translation. These results suggest that with proper ASR and MT adaptation through in-domain data, we could obtain similar English-speaking IA performance on machine translation outputs as the Spanish-speaking IAs on

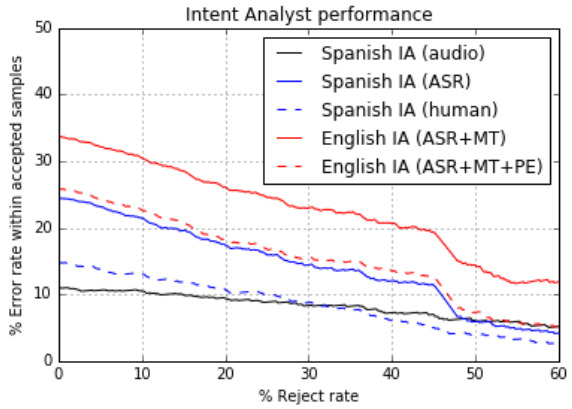

Figure 3: Error-rejection curves for the intent analyst (IA) labeling accuracy on Spanish audio, Spanish ASR and human transcripts, and their machine translations into English (MT). Rejection is computed from the English model's confidence scores on MT outputs. "+PE": performance on post-edited MT outputs.

their native language utterances.

9 Related Work

The use of MT to translate texts in other languages into English for sentiment analysis was proposed in Denecke (2008). Bautin et al. (2008) show that sometimes MT performs inadequate translations on essential parts of a text, affecting sentiment analysis performance. Our results confirm this phenomena due to a lack of in-domain MT training data. Schwenk and Douze (2017) explore learning multilingual sentence embeddings with neural MT, which can aid in multilingual search. Prior to that, multilingual approaches leveraged lexical resources such as MultiWordNet (Pianta et al., 2002) to bridge concepts from one language to another.

10 Conclusions

We have executed an experiment to measure machine translation's ability to rapidly bootstrap intent classification models for new languages. In our English→Spanish experiments, we observe that although the initial results appear to be substantially worse than a Native Spanish intent classification model, we show that MT can provide a degree of automation that supports human-assisted multilingual dialog systems that can be deployed to production on day one, reducing the need for human agent support over a fully manual solution. There is further promise that model improvements can be obtained by improving the ASR and machine translation models to include in-domain data. Finally, we observe it is better to use the online Spanish→English bootstrap in our production system rather than an offline English→Spanish intent model.

References

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural Machine Translation by Jointly Learning to Align and Translate. In *5th International Conference on Learning Representations*, San Diego, USA. ICLR.

Bautin, Mikhail, Lohit Vijayarenu, and Steven Skiena. 2008. International sentiment analysis for news and blogs. In Adar, Eytan, Matthew Hurst, Tim Finin, Natalie S. Glance, Nicolas Nicolov, and Belle L. Tseng, editors, *ICWSM*. The AAAI Press.

Bojar, Ondřej, Christian Buck, Chris Callison-Burch, Christian Federmann, Barry Haddow, Philipp Koehn, Christof Monz, Matt Post, Radu Soricut, and Lucia Specia. 2013. Findings of the 2013 Workshop on Statistical Machine Translation. In *Proceedings of the Eighth Workshop on Statistical Machine Translation*, pages 1–44, Sofia, Bulgaria, August. Association for Computational Linguistics.

Cettolo, Mauro, Jan Niehues, Sebastian Stüker, Luisa Bentivogli, and Marcello Federico. 2014. Report on the 11th IWSLT Evaluation Campaign. In *Proceedings of the International Workshop on Spoken Language Trnaslation (IWSLT)*, Lake Tahoe, USA, December.

Denecke, K. 2008. Using sentiwordnet for multilingual sentiment analysis. In *2008 IEEE 24th International Conference on Data Engineering Workshop*, pages 507–512, April.

Luong, Minh-Thang and Christopher D. Manning. 2015. Stanford neural machine translation systems for spoken language domain. In *International Workshop on Spoken Language Translation*, Da Nang, Vietnam.

Pianta, Emanuele, Luisa Bentivogli, and Christian Girardi. 2002. Multiwordnet: developing an aligned multilingual database. In *Proceedings of the First International Conference on Global WordNet*, January.

Schwenk, Holger and Matthijs Douze. 2017. Learning joint multilingual sentence representations with neural machine translation. In *Proceedings of the 2nd Workshop on Representation Learning for NLP*, pages 157–167. Association for Computational Linguistics.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016, August 7-12, 2016, Berlin, Germany, Volume 1: Long Papers*.