INTERSPEECH 2009 BRIGHTON

Localization of Speech Recognition in Spoken Dialog Systems: How Machine Translation Can Make Our Lives Easier

David Suendermann, Jackson Liscombe, Krishna Dayanidhi, Roberto Pieraccini

SpeechCycle Labs, New York, USA

{david, jackson, krishna, roberto}@speechcycle.com

10.21437/Interspeech.2009-450

Abstract

The localization of speech recognition for large-scale spoken dialog systems can be a tremendous exercise. Usually, all involved grammars have to be translated by a language expert, and new data has to be collected, transcribed, and annotated for statistical utterance classifiers resulting in a time-consuming and expensive undertaking. Often though, a vast number of transcribed and annotated utterances exists for the source language. In this paper, we propose to use such data and translate it into the target language using machine translation. The translated utterances and their associated (original) annotations are then used to train statistical grammars for all contexts of the target system. As an example, we localize an English spoken dialog system for Internet troubleshooting to Spanish by translating more than 4 million source utterances without any human intervention. In an application of the localized system to more than 10,000 utterances collected on a similar Spanish Internet troubleshooting system, we show that the overall accuracy was only 5.7% worse than that of the English source system.

Index Terms: spoken dialog systems, machine translation, localization

1. Introduction

Nowadays' spoken dialog systems can be very complex applications comprising thousands of activities, grammars, and prompts. Years of developing work can be spent to design these systems and much effort undertaken to tune involved speech recognition grammars to achieve highest possible performance crucial for user acceptance and effectiveness of the applications. Such tuning can require processing of huge numbers of calls to analyze caller behavior in every single context of the system, building of recognition grammars to effectively interpret caller utterances, and designing the application to respond appropriately at every context.

E.g., to tune a spoken dialog system for Internet, cable TV, and Voice-over-IP troubleshooting, more than two million speech utterances were collected, transcribed, annotated, and used for training statistical grammars, boosting overall accuracy from an initial 78.0% to 90.5% accuracy [1]. Although transcription and annotation of such amounts of data is partially automatable, it can still keep several people busy for months. While transcription is a relatively straightforward exercise, semantic annotation, i.e. the mapping of a lexical content to one of a number of semantic symptoms, requires knowledge about the application. Not only must annotators understand what a caller utterance means in response to the system prompt in the respective context, but there are several aspects to semantic annotation making it a non-trivial undertaking, such as

- Utterances may have no representation in the given set of symptoms suggesting that they are out-of-scope for the grammar.
- When the ratio of out-of-scope utterances grows and well-distinguishable patterns manifest themselves, annotators are to suggest the introduction of new symptoms to the system designer.
- Utterances may be ambiguous, vague, too specific, or carry content belonging to multiple symptoms making it hard for the annotator to make a decision.
- Annotations have to follow a number of quality assurance criteria to produce powerful and exact results including criteria for completeness, consistency, congruence, correlation, confusion, coverage, and corpus size, also referred to as C^7 [2].

These issues emphasize that thorough speech recognition tuning in spoken dialog systems can be a very expensive task. Large-scale spoken dialog systems as introduced above are mostly used in relatively big enterprises trying to optimize their customer care telephone portals. Many of these companies operate internationally producing a need to localize their phone services including involved spoken dialog systems. Localization of a dialog system entails translating it from one language to another [3]. The high cost of producing and maintaining systems in different languages obviously increases as more languages are considered. Not only the cost, but also the time to generate speech recognition grammars from scratch is a crucial issue when localizing a given spoken dialog system [4].

In this paper, we propose to use transcribed and annotated data available for the original (source) language of the spoken dialog system, then apply machine translation to the given transcriptions keeping the semantic annotations, and finally training statistical grammars based on the translated utterances and the original annotations. As a proof of concept, we used all available data collected for an English Internet troubleshooting application comprising more than 4 million utterances to build

Patent pending.

Spanish grammars for every recognition context of the application. Then we collected, transcribed, and annotated 951 full calls (11470 utterances) of a Spanish Internet troubleshooting application. Testing the utterances against the translated statistical grammars for the given recognition contexts resulted in an average accuracy of 85.0%. As a comparison, the original English system performed at 90.7%.

2. Some Theoretical Background

The subject discussed in this paper, the localization of speech recognition based on machine translation, is related to several areas of speech processing including automatic speech recognition, machine translation, and speech translation. This section is to give a very high-level overview on the main probabilistic apparatus of these related disciplines to indicate how they are mathematically interconnected.

2.1. Speech Recognition (Speech F to Text f)

In the digital age, the usual input to speech recognition is a pulse-code modulated (or similarly coded) chunk of audio which most often is transformed to a sequence of feature vectors F. Given this vector sequence, the non-trivial problem is to find the most probable sequence of words

$$f = \arg\max_{\varphi} p(\varphi|F) \tag{1}$$

where φ iterates over the set of all possible word sequences. Bayes' theorem allows to rewrite this formula into

$$f = \arg\max_{\varphi} p(\varphi)p(F|\varphi). \tag{2}$$

Here, $p(\varphi)$ is the probability of the word sequence φ, commonly referred to as language model, whereas $p(F|\varphi)$ is the conditional probablity that the feature vector sequence F was produced by the word sequence φ, referred to as acoustic model [5].

2.2. Machine Translation (Text f to Text e)

Machine translation can be described similarly by searching for that word sequence of the target language e being the most likely translation of the source word sequence f:

$$e = \arg\max_{\varepsilon} p(\varepsilon|f) \tag{3}$$

where ε iterates over the set of all possible target word sequences. We can apply Bayes' theorem producing

$$e = \arg\max_{\varepsilon} p(\varepsilon)p(f|\varepsilon) \tag{4}$$

with the target language model $p(\varepsilon)$ and the so-called translation model $p(f|\varepsilon)$ which expresses the probability that the source (or foreign) language word sequence f is the translation of the target (or native) language word sequence ε. This somewhat counter-intuitive splitting of the problem into two subproblems where the second one (the translation model) looks as hard as the original problem is motivated by the fact that the first subproblem (the language model) has a significant impact for the success of the search expressed by Equation 4 and can be relatively straightforwardly be estimated based on large amounts of target language data [6].

2.3. Speech Translation (Speech F to Text e)

The coupling of automatic speech recognition and machine translation—actually out-of-scope of the present paper but included in this section for the sake of completeness—allows for directly translating spoken utterances into another language [7]. Here, we search for the most probable target language word sequence e given an acoustic source vector sequence F as

$$\begin{aligned} e &= \arg\max_{\varepsilon} p(\varepsilon|F) \\ &= \arg\max_{\varepsilon} p(\varepsilon)p(F|\varepsilon) \\ &= \arg\max_{\varepsilon} p(\varepsilon) \sum_{\varphi} p(F|\varphi,\varepsilon)p(\varphi|\varepsilon) \\ &\cong \arg\max_{\varepsilon} p(\varepsilon) \sum_{\varphi} p(F|\varphi)p(\varphi|\varepsilon). \end{aligned} \tag{5}$$

The last step's approximation assumes that the acoustic realization of an utterance in a language only depends on the underlying word sequence of the same language and is independent of its translation into another language. Here, we find target language model, source acoustic model, and translation model in combination.

2.4. Speech Recognition Localization (Speech E to Text e)

Now, we want to localize speech recognition to another (a target) language, i.e., we want to transcribe the feature sequence E to a word string e as per Equation 2:

$$e = \arg\max_{\varepsilon} p(E|\varepsilon)p(\varepsilon). \tag{6}$$

Mostly, in commercial applications, the acoustic model $p(E|\varepsilon)$ is provided by the speech recognizer's manufacturer whereas the target language model (in spoken dialog systems aka grammars) will most often be context- and application-dependent, i.e., it has to be rebuilt. According to Section 1, we propose to apply knowledge we have from the source language as can be expressed by extending Equation 6 as follows:

$$e = \arg\max_{\varepsilon} p(E|\varepsilon) \sum_{\varphi} p(\varphi)p(\varepsilon|\varphi). \tag{7}$$

This formulation leaves us with the translation model $p(\varepsilon|\varphi)$ implemented in a machine translation environment as discussed in Section 2.2 as well as with the source language model $p(\varphi)$ whose approximation produces no additional cost in the present localization scenario due to the large set of source utterances available.

3. The Source Data

As mentioned in Section 1, as an example case, we used source data collected in the scope of a large-scale English dialog system for broadband Internet troubleshooting as described in further detail in [8]. Over a time span of more than three years, dozens of millions of calls were processed by this system. On a subset of these calls, utterances were captured, transcribed, and annotated according to their semantic meaning. Table 1 gives an overview about the amount of involved data listing the number of calls with transcribed utterances, the number of transcribed and annotated utterances, activities, and grammars. Due to a continuous improvement cycle applied to the example application, several existing grammars were regularly updated by

Table 1: *Overview on the English source data.*

calls	1,159,940
transcribed utterances	4,293,898
annotated utterances	3,846,050 (89.6%)
activities	2,332
grammars	253
root grammars	134

optimized statistical language models and classifiers [1]. Consequently, several versions of grammars in the same recognition context were used over the time of the data collection. Since for the purpose of the present exercise all the data collected in such contexts is to be used independently of the actual grammar version active at the time of the utterance capture, we do not distinguish between contexts originating from the same original or *root* grammar. Also the number of root grammars is given in Table 1.

Figure 1 shows the distribution of these utterances over the mentioned time period indicating that the capture volume was ever-increasing since the start of the project.

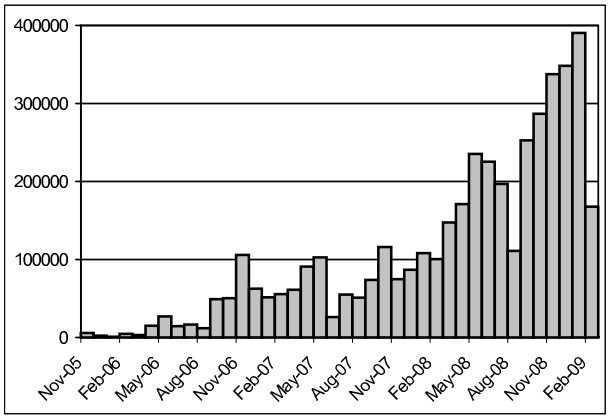

Figure 1: *Utterance volume per month.*

4. The Experiment

4.1. Translation

All transcribed utterances of Table 1 were translated from English into Spanish using a commercial statistical machine translation software. This was done completely unsupervised. No corrections of the output or any tuning of the machine translator was performed.

4.2. Training

For all distinct root grammars of Table 1, the respective translated Spanish utterances and their original semantic annotations were used to train a statistical language model and a statistical classifier using standard settings for the involved parameters, since no development data was available[1]. These settings are given in Table 2. Figure 2 shows the (Zipf-like) distribution

[1] Development data would have to be based on Spanish speech data since language model and classifier have to be applied to a speech recognizer in the target language.

Table 2: *Training settings.*

language model	trigram + smoothing
classifier	naïve Bayes + boosting
language/acoustic model tradeoff	0.8
training accuracy cutoff	99%
acoustic rejection threshold	5%
semantic rejection threshold	0%

Table 3: *Overview on the Spanish testing data.*

calls	951
transcribed utterances	11,470
annotated utterances	11,470 (100.0%)
activities	144
grammars	17

of the number of utterances for each of the grammars in descending order showing that there are grammars exceeding one million utterances (a typical yes/no context) as well as numerous grammars facing data sparseness (22 grammars feature less than 100 training utterances).

4.3. Test

To test (a subset of) the automatically translated grammars, we collected, transcribed, and annotated a limited number of utterances from a Spanish version of a similar broadband Internet troubleshooting dialog system. The characteristics of this data are shown in Table 3. Figure 2 indicates the grammars found in the test data as white bullets showing that they are distributed among different magnitutes of amounts of available training data.

Now, a batch experiment was executed performing speech recognition and classification on the complete set of collected utterances using the automatically translated grammars in their respective contexts. For each of the 11,470 utterances, the classification result was now compared to the semantic annotation of the same utterance. In the following, we refer to accuracy as the number of acoustic events where classification result and annotation match divided by the total number of acoustic events. These events include out-of-scope utterances as well as noise,

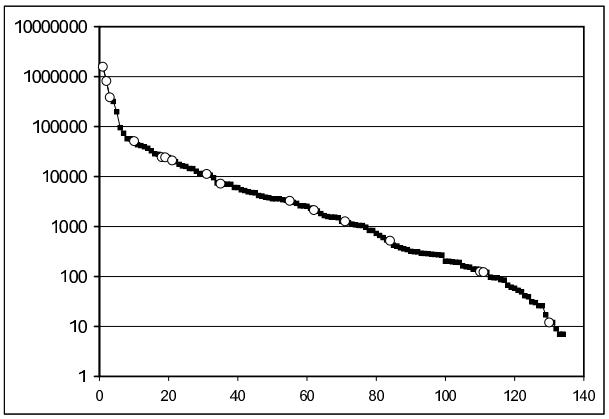

Figure 2: *Number of utterances per root grammar in order of descending frequency.*

background speech, etc.

Overall accuracy for the entire test set was at 85.0% which is deemed very high compared to the performance of most boot-strapped dialog systems based on hand-crafted grammars. Those systems often perform at less than 80% accuracy; an example is given in Section 1. To have a more reliable standard of comparison, we looked at the performance of the English source dialog system optimized on performance for several years and found that the latest available system version performed at 90.7% (measured on 930 full calls, 11274 completely annotated utterances).

5. Conclusion

We have shown that localizing speech recognition using machine translation can be straightforward and cheap when large amounts of transcribed and annotated data of the source language is available. Testing an example implementation of the proposed methodology indicated that this approach outperforms manual boot-strapping but does not achieve the same accuracy like the original (source language) dialog system. The reason for the performance loss can be explained by the weakness of either of the factors in Equation 7:

- The target acoustic model $p(E|\varepsilon)$ is weak: In our experiment, we used an out-dated Spanish speech recognizer whose acoustic models obviously did not achieve the same performance like its English couterparts. E.g., in yes/no (sí/no) contexts, we saw a significantly higher portion of false accepts and rejects than in equivalent English contexts clearly independent of any linguistic factors.
- The translation model $p(\varepsilon|\varphi)$ is weak: Statistical translation not only produces a lot of commonly known artifacts, but there are cases where even a human translator would fail: A grammar is normally designed based on utterances a caller says in response to a system prompt restricting the caller's language. For instance, a Spanish prompt may say

  "cuando esté desconectado, diga *continúe*"

  translated from the English prompt

  "when it's unplugged, say *continue*".

  Hence, most of the English reponses will read "continue" which a machine as well as a human being most likely would translate into Spanish as "continuar" instead of the prompt-dependent correct "continúe". So, to achieve a higher accuracy of the translation hypotheses, they could be rescored taking the respective system prompt and other application-dependent information into consideration.

Furthermore, as mentioned in Section 4.1, no development data was available for this experiment since this would have required a (minimal) portion of collected target language utterances, their transcriptions and annotations. Such data may be available once the first version of the target system goes into production and can be used to tune language models and classifiers.

6. References

[1] D. Suendermann, J. Liscombe, K. Evanini, K. Dayanidhi, and R. Pieraccini, "From Rule-Based to Statistical Grammars: Continuous Improvement of Large-Scale Spoken Dialog Systems," in *Proc. of the ICASSP*, Taipei, Taiwan, 2009.

[2] D. Suendermann, J. Liscombe, K. Evanini, K. Dayanidhi, and R. Pieraccini, "C^5," in *Proc. of the SLT*, Goa, India, 2008.

[3] H. Strik, A. Russel, H. v. d. Heuvel, C. Cucchiarini, and L. Boves, "Localizing an Automatic Inquiry System for Public Transport Information," in *Proc. of the ICSLP*, Philadelphia, USA, 1996.

[4] N. Perera and A. Ranta, "Dialogue System Localization with the GF Resource Grammar Library," in *Proc. of the ACL Workshop on Grammar-Based Approaches to Spoken Language Processing*, Prague, Czech Republic, 2007.

[5] L. Rabiner and B. Juang, *Fundamentals of Speech Recognition*, Prentice Hall, Englewood Cliffs, USA, 1993.

[6] F. Och, "Challenges in Machine Translation," in *Proc. of the TC-Star Workshop*, Barcelona, Spain, 2006.

[7] E. Matusov, S. Kanthak, and H. Ney, "On the Integration of Speech Recognition and Statistical Machine Translation," in *Proc. of the Interspeech*, Lisbon, Portugal, 2005.

[8] K. Acomb, J. Bloom, K. Dayanidhi, P. Hunter, P. Krogh, E. Levin, and R. Pieraccini, "Technical Support Dialog Systems: Issues, Problems, and Solutions," in *Proc. of the Workshop on Bridging the Gap: Academic and Industrial Research in Dialog Technologies*, Rochester, USA, 2007.