MT and Topic-Based Techniques to Enhance Speech Recognition Systems for Professional Translators

Yevgeny Ludovik and Ron Zacharski
Computing Research Laboratory
New Mexico State University
Las Cruces, New Mexico
{eugene, raz}@crl.nmsu.edu

Abstract

Our principle objective was to reduce the error rate of speech recognition systems used by professional translators. Our work concentrated on Spanish-to-English translation. In a baseline study we estimated the error rate of an off-the-shelf recognizer to be 9.98%. In this paper we describe two independent methods of improving speech recognizers: a machine translation (MT) method and a topic-based one. An evaluation of the MT method suggests that the vocabulary used for recognition cannot be completely restricted to the set of translations produced by the MT system and a more sophisticated constraint system must be used. An evaluation of the topic-based method showed significant error rate reduction, to 5.07%.

Introduction

Our goal is to improve the throughput of professional translators by using speech recognition. The problem with using current off-the-shelf speech recognition systems is that these systems have high error rates for similar tasks. If the task is simply to recognize the speech of a person reading out loud, the error rate is relatively low; the error rate of large vocabulary research systems (20,000-60,000 word vocabularies) performing such a task is, at best, around 10% (see, for example, Robinson and Christie 1998, Renals and Hochberg 1996, Hochberg et al. 1995 and Siegler and Stern 1995). The popular press has reported slightly lower results for commercial systems. For example, PC Magazine (Poor 1998) compared Dragon's NaturallySpeaking and IBM's ViaVoice (both continuous speech recognition systems with approximately 20,000 word vocabularies). They evaluated these systems by having five speakers read a 350 word text at a slow pace (1.2 words/second) after completing a half hour training session with each system. The average recognition error rate was 11.5% (about 40 errors in the 350 word text). An evaluation of the same two systems without training resulted in a recognition error rate of 34% (Keizer 1998). If the task is more difficult than recognizing the speech of a person reading, the error rate increases dramatically. For example, Ringger (1995) reports an average error rate of 30% for recognizing careful, spontaneous speech on a specific topic. However, the error rate of paced speech can be as low as 5% if the vocabulary is severely limited or if the text is highly predictable and the system is tuned to that particular genre. Unfortunately, the speech of expert translators producing spoken translations does not fall into any of the "easy to recognize" categories.

In many translation tasks the source document is in electronic form and the obvious question to ask is if an analysis of the source document could lead to a reduction of the speech recognition error rate. For example, suppose we have a robust machine translation system and use it to generate all the possible translations of a given source text. We could then use this set of translations to help predict what the translator is saying. We describe this approach in §1 below. A simpler approach is to identify the topic of the source text and use that topic to aid in speech recognition. Such as approach is described in §2. Both methods were tested in a Spanish-to-English translation task.

This research rests on two crucial ideas. The first is that lexical and translation knowledge extracted from source documents by automated natural language processing can be utilized in a large-vocabulary, continuous speech recognizer to achieve low word-error rates. The second idea is that the translator should be able to dictate a translation and correct the resulting transcription in much less time than if they had to type the translation themselves or rely on a transcriber/typist.

1. Using machine translation

The difference between a typical speech dictation system and the situation described above, is that the translator is viewing the source text on a computer—that is, the text is available online. This source text can be analyzed using a machine translation (MT) component. Hopefully, this analysis will cut down on the recognition perplexity by having the recognizer make choices only from the set of possible renderings in the target language of the words in the source language. In this section we describe the MT subsystem in detail.

The function of this subsystem is to take Spanish sentences as input and produce a set of English words that are likely to occur in translations of these sentences. For example, if the Spanish text is

1. *Butros Ghali propone vía diplomática para solucionar crisis haitiana*

we would expect the translation set to include the words (among others):

{*Boutros, Ghali, proposes, diplomatic, route, to, settle, Haitian, crisis*}

Hopefully, this translation set will be a good predictor of what the translator actually said.

1.1 The MT subsystem

The MT subsystem consists of 4 components: the Spanish morphological analyzer, the dictionary lookup component, the lexical transfer component, and the English morphological generator. These components are briefly described in this section.

1.1.1 Spanish morphological analyzer

The morphology analyzer takes Spanish words as input and outputs a set of possible morphological analyses for those words. Each analysis consists of the root word and a set of feature structures representing the information obtained from inflectional morphology. Examples are given below.

Word	**Feature structure**
cafés	((root café) (cat n) (number plural))
pequeña	((root pequeño)(cat adj)(gender f))
podría	((root podrir) (cat v) (tense imperfect indicative) (person 3)(number singular))

1.1.2 Dictionaries and dictionary lookup

The dictionary lookup component takes a feature structure produced by the morphological analyzer, looks up the root-word/part-of-speech pair in the dictionary, and adds information to the existing feature structure. The words in the dictionary were derived from doing a corpus analysis of a set of 20 Spanish test documents. All the unique words in this corpus, including proper nouns, were included in the dictionary (approximately 1,500 words). A few examples are shown below.

actividad	((root actividad) (cat n) (trans activity energy) (gender f))
comenzar	((root comenzar)(cat v)(trans begin start) (verbtype irregular 129))
cuestion	((root cuestion) (cat n) (trans question dispute problem issue) (gender f))

1.1.3 The lexical transfer component

At the end of the dictionary lookup phase, for each word in the Spanish sentence we have a feature structure containing the information in the dictionary entry along with the parameter values that were gained from morphological analysis. One feature, trans, contains the possible English translations of that Spanish word. The lexical transfer component converts this Spanish feature structure to one or more English feature structures; one feature structure is created for each value in the trans field. For example, the feature structure associated with an instance of *actividad* encountered in some text

will be 'transferred' to two English feature structures: one for *activity* and one for *energy*. Similarly, encountering a *cuestion* in some text, will result in the creation of four feature structures; those representing the English words *question*, *dispute*, *problem*, and *issue*. In addition, the transfer component converts other features in the Spanish feature structure to features recognizable to the English morphological generator.

1.1.4 The English morphological generator

We used an English Morphological generator developed at the Computing Research Laboratory at New Mexico State University by Steve Beale. The morphological generator takes feature structures as input and produces correctly inflected English words. Examples of the feature structures used as input and their associated output are illustrated below:

((root run) (cat v) (num plural)(form progressive))	*are running*
((root run) (cat v) (tense future) (form progressive))	*will be running*
((root man) (cat n) (number plural))	*men*

1.2 Evaluation

Suppose we wish to have a user dictate an English translation of a Spanish sentence that appears on a computer screen. This Spanish sentence is input to the MT system and the output is a set of English words. In the ideal case, the words in the English sentence the translator dictates are contained in this set. If one could offer a sort of guarantee that the words of any reasonable translation of the Spanish sentence are contained within this set, then incorporating the MT subsystem into a speech recognition system would be relatively straight forward; the vocabulary at any given moment would be restricted to this word set. If, on the other hand, such a guarantee cannot be made then this approach will not work. The evaluation of the natural language subsystem is designed to test whether reasonable translations are contained within this set of words.

The test material consisted of 10 Spanish newspaper articles. The articles were translated into English by two independent translators. The following table shows that roughly 1/3 of the words in the translations the professional translators produced are not in the set of words produced by the natural language subsystem (T1 and T2 are the two different English translations):

Table 1 : % of words in translation not in word set

Document number	T 1	T 2
1	30.4	26.78
2	30.08	33.16
3	37.88	32.66
4	32.03	39.21
5	27.69	23.79
6	31.3	27.79
7	32.85	30.25
8	34.84	31.32
9	43.8	40.05
10	34.95	34.5

Average: 32.77

The next experiment augmented the word set constructed by the approach described above with the 800 most frequent words in a 2 million word corpus of English. The results are illustrated in the following table.

Table 2 : % of words in translation that are not in the word set: frequent wordlist & morphological analysis

1	12.7	14.21
2	16.89	15.05
3	19.22	18.62
4	10.68	16.05
5	13.85	12.53
6	13.33	12.39
7	15.41	14.01
8	19.1	16.38
9	17.47	15.25
10	19.42	16.61

Average: 15.46%

The reason this combined method was tested was that often English open class lexical items are added to the translation. For example in one document, the phrase *solucionar crisis haitiana* is translated as "resolution of Haitian crisis", and the English *of* does not have a direct correlate in the Spanish phrase. While this combined method appears to work moderately well, it still does not have sufficient coverage to function as a method for generating the

complete recognition vocabulary. That is, it cannot guarantee that the words of any reasonable translation of a Spanish sentence would be contained in the set of English words generated from that sentence. Since we cannot use an MT system to constrain the recognition vocabulary we evaluated a different method—one that uses topic recognition.

2. Topic recognition method

The basic idea behind the topic recognition approach is to identify the topic of the source language text and then use that topic to alter the language model for speech recognition.

2.1 Topic recognition of source text

We used a naïve Bayes classifier to identify the topic of Spanish online newspaper texts. We eliminated the common words in the text under the rubric that these words are unlikely to serve as cues to the topic of the text. For example in English, *the*, *of*, *with*, and *a* provide little information as to the topic of the text. We constructed this common word list by computing the most frequent words in a one million word corpus of Spanish newspaper text. This list was edited to remove potential topic cues. For example, *Pinochet* was the 46th most frequent word and *Clinton* was the 65th most frequent, but they serve as potential topic cues. We evaluated this topic identification technique by examining its performance on identifying four topics: Pinochet, the crisis in Paraguay, the crisis in Kosovo, and Clinton's impeachment. For each topic we had a 500k training corpus (roughly 60,000-75,000 words). The test data for each topic consisted of 20 articles from web-based newspapers. The average size of these articles was 335 words. The recognition results are shown in the following table:

Table 3 : Accuracy of topic recognition

Words used in recognition	Pinochet	Paraguay	Kosovo	Clinton
all	100	100	100	100
100	100	100	95	100
50	95	100	95	100
25	90	95	90	95

We also evaluated an enhanced version of the algorithm on a corpus of 20 newsgroups.[1] For this evaluation we used a different method of creating a common word list. For each work encountered in any training document we computed the entropy of the distribution of a topic given the word, and picked up 100 words having the highest entropy. No manual editing of this list was done. High entropy for a given word meant that this word could not be a good topic cue. In this evaluation for each value of the number of words used in recognition we carried out two sets of experiments. In the first, the first 500 documents of each topic were used as training data, and the last 500 as test data; in the second, the last 500 documents were used as training data and the first 500 as test. The recognitioin results are presented in the following table.

Table 4 : Topic recognition results for 20 newsgroups : 100 common words excluded

Words used in recognition	Recognition rate
all	76.76
100	53.15
50	48.41
25	44.23

2.2 Using topic language models

In the previous section we have described a robust topic recognition system and describe how the system performed in identifying the topic of Spanish texts. Once we have identified the topic of the text to be translated we use that topic to identify which language models we wish to use in recognizing the text. We have constructed topic language models using IBM's ViaVoice Topic Factory, which allows allows developers to construct specialized language models that augment the main recognition language model. To construct these models we manually collected half million word corpora for both the crisis in Kosovo and Clinton's impeachment. These corpora were collected from a variety of online news sites including

[1] from Tom M. Mitchell's website http://www.cs.cmu.edu/afs/cs/project/theo-11/www/naive-bayes.html

CNN, the Washington Post, the New York Times, the New York Daily News, and the Milwaukee Journal Sentinel. One significant question is whether a language model as small as a half a million words will have any impact on the error rate for speech recognition. We evaluated this approach by comparing the error rate in dictating 8 texts. The results are shown in the table below. (The 'without' row is using the recognizer without our topic system and the 'with' row uses it with topic identification.)

Table 4: Dictation error rates

text #	without	with
1	8.59	5.62
2	8.67	6.15
3	10.16	4.46
4	8.88	4.75
5	12.07	5.26
6	13.47	6.15
7	8.17	4.93
8	9.8	3.27
average	9.98	5.07

As this table shows the topic-based method reduces the average error rate by approximately 49%. This is rather remarkable given the simplicity of the method and the extremely small training corpus for the language model.

Conclusion

In this paper we reviewed two methods for reducing speech recognition errors rates. The first method used a word-for-word MT system to constrain recognition vocabulary. Results of an evaluation of this method suggest that an MT system cannot adequately predict what words will be used in an actual translation and a more sophisticated method of incorporating MT into a recognizer is needed. For example, we could extend our MT system to construct a set of possible translations for the entire source language sentence. We could then use this set of English sentences to train a small language model, which would be used to recognize the sentences the translator produced. Alternatively, we could use a translation memory approach to MT to construct the set of English sentences (Webb 1992). The second method we described recognized the topic of the source document and used a language model associated with that topic for speech recognition. Using this approach, the error rate was reduced from 9.98 to 5.07%. This means, for example, that for a short, 1 page, 500 word document, this method has saved the translator the time it would take to go back and manually correct 25 errors.

Acknowledgements

This work was partially funded by NSF grant DMI-9860308 to Onyx Consulting, Inc. in Las Cruces, New Mexico. We would like to thank Sergei Nirenburg and Jim Cowie for their assistance.

References

Hochberg, M. Renals, S., Robinson, A. and Cook, G. (1995) Recent improvements to the Abbot Large Vocabulary CSR System. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing, 69-72.

Keizer, Gregg. 1998. The gift of gab: CNET compares the top speech recognition apps. (http://204.162.80.182/Content/Reviews/Compare/Speech/).

Poor, Richard. 1998. Speech recognition: watch what you say. PC Magazine on-line (http://home.zdnet.com/pcmag/features/speech/index.html).

Renals, S and Hochberg, M. (1996) Efficient evaluation of the LVCSR search space using the NOWAY decoder. Proceedings of the International Conference on Speech and Language Processing, 149-152.

Ringger, Eric K. (1995) A robust loose coupling for speech recognition and natural language understanding. Technical Report 592. University of Rochester Computer Science Department.

Robinson, T. and Christie, J. (1998) Time-first search for large vocabulary speech recognition. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Siegler, M. and Stern R. (1995) On the effects of speech rate in large vocabulary speech recognition systems. Proceedings of the International Conference on Acoustics, Speech, and Signal Processing.

Webb, L. (1992) Advantages and disadvantages of translation memory: a cost/benefit analysis. Monterey Institute of International Studies MA Thesis.