Beyond N-Grams: Can Linguistic Sophistication Improve Language Modeling?

Eric Brill, Radu Florian, John C. Henderson, Lidia Mangu
Department of Computer Science
Johns Hopkins University
Baltimore, Md. 21218 USA
{brill,rflorian,jhndrsn,lidia}@cs.jhu.edu

Abstract

It seems obvious that a successful model of natural language would incorporate a great deal of both linguistic and world knowledge. Interestingly, state of the art language models for speech recognition are based on a very crude linguistic model, namely conditioning the probability of a word on a small fixed number of preceding words. Despite many attempts to incorporate more sophisticated information into the models, the n-gram model remains the state of the art, used in virtually all speech recognition systems. In this paper we address the question of whether there is hope in improving language modeling by incorporating more sophisticated linguistic and world knowledge, or whether the n-grams are already capturing the majority of the information that can be employed.

Introduction

N-gram language models are very crude linguistic models that attempt to capture the constraints of language by simply conditioning the probability of a word on a small fixed number of predecessors. It is rather frustrating to language engineers that the n-gram model is the workhorse of virtually every speech recognition system. Over the years, there have been many attempts to improve language models by utilizing linguistic information, but these methods have not been able to achieve significant improvements over the n-gram.

The insufficiency of Markov models has been known for many years (see Chomsky (1956)). It is easy to construct examples where a trigram model fails and a more sophisticated model could succeed. For instance, in the sentence : *The dog on the hill **barked***, the word **barked** would be assigned a low probability by a trigram model. However, a linguistic model could determine that **dog** is the head of the noun phrase preceding **barked** and therefore assign **barked** a high probability, since P(barked|dog) is high.

Using different sources of rich linguistic information will help speech recognition if the phenomena they capture are prevalent and they involve instances where the recognizer makes errors.[1] In this paper we first give a brief overview of some recent attempts at incorporating linguistic information into language models. Then we discuss experiments which give some insight into what aspects of language hold most promise for improving the accuracy of speech recognizers.

1 Linguistically-Based Models

There is a continuing push among members of the speech recognition community to remedy the weaknesses of linguistically impoverished n-gram language models. It is widely believed that incorporating linguistic concepts can lead to more accurate language models and more accurate speech recongizers.

One of the first attempts at linguistically-based modelling used probabilistic context-free grammars (PCFGs) directly to

[1] This is one of the problems with perplexity as a measure of language model quality: if the *better* model simply assigns higher probability to the elements the recognizer already gets correct, the model will look better in terms of perplexity, but will do nothing to improve recognizer accuracy.

compute language modeling probabilities (Jelinek(1992)). Another approach retrieved n-gram statistics from a handwritten PCFG and combined those statistics with traditional n-grams elicited from a corpus (Jurafsky(1995)). Research has been carried out in adaptively modifying language models using knowledge of the subject matter being discussed (Seymore(1997)). This research depends on the prevalence of jargon and domain-specific language.

Linguistically motivated language models were investigated for two consecutive years at the Summer Speech Recognition Workshop, held at Johns Hopkins University. In 1995 experiments were run adding part-of-speech (POS) tags to the language models (Brill(1996)). In the 1996 Summer Speech Recognition Workshop, recognizer improvements were attempted by exploiting the long-distance dependencies provided by a dependency parse (Chelba(1997)). The goal was to exploit the predictive power of predicate-argument structures found in parse trees. In Della Pietra(1994) and Fong(1995), link grammars were used, again in an attempt to improve the language model by providing it with long-distance dependencies not captured in the n-gram statistics.[2]

Although much work has been done exploring how to create linguistically-based language models, improvement in speech recognizer accuracy has been elusive.

2 Experimental Framework

In an attempt to gain insight into what linguistic knowledge we should be exploring to improve language models for speech recognition, we ran experiments where people tried to improve the output of speech recognition systems and then recorded what types of knowledge they used in doing so. We hoped to both assess how much gain might be expected from very sophisticated models and to determine just what information sources could contribute to this gain.

People were given the ordered list of the ten most likely hypotheses for an utterance according to the recognizer. They were then asked to choose from the ten-best list the hypothesis that they thought would have the lowest word error rate, in other words, to try to determine which hypothesis is closest to the truth. Often, the truth is not present in the 10-best list. An example 5-best list from the Wall Street Journal corpus is shown in Figure 1. Four subjects were used in this experiment, and each subject was presented with 75 10-best lists from three different speech recognition systems (225 instances total per subject). From this experiment, we hoped to gauge what the upper bound is on how much we could improve upon state of the art by using very rich models.[3]

For our experiments, we used three different speech recognizers, trained respectively on Switchboard (spontaneous speech), Broadcast News (recorded news broadcasts) and Wall Street Journal data.[4] The word error rates of the recognizers for each corpus are shown in the first line of Table 1.

The human subjects were presented with the ten-best lists. Sentences within each ten-best list were aligned to make it easier to compare them. In addition to choosing the most appropriate selection from the 10-best list, subjects were also allowed to posit a string not in the list by editing any of the strings in the 10-best list in any way they chose. For each sample, subjects were asked to determine what types of information were used in deciding. This was done by presenting the subjects with a set of check boxes, and asking them to check all that applied. A list of the options presented to the human can be found in Figure 2. Subjects were provided with a detailed explanation, as well as examples, for each of these options.[5]

2 Net Human Improvement

The first question to ask is whether people are able to improve upon the speech recognizer's output by postprocessing the n-best lists. For

[2] For a more comprehensive review of the historical involvement of natural language parsing in language modelling, see Stolcke(1997).

[3] Note that what we are really measuring is an upper bound on improvement under the paradigm of n-best postprocessing. This is a common technique in speech recognition, but it results in the postprocessor not having access to the entire set of hypotheses, or to full acoustic information.

[4] HTK software was used to build all recognizers.

[5] This program is available at http://www.cs.jhu.edu/labs/nlp

each corpus, we have four measures: (1) the recognizer's word error rate, (2) the oracle error rate, (3) human error rate when choosing among the 10-best (human selection) and (4) human error rate when allowed to posit any word sequence (human edit).

The oracle error rate is the upper bound on how well anybody could do when restricted to choosing between the 10 best hypotheses: the oracle always chooses the string with the lowest word error rate. Note that if the human always picked the highest-ranking hypothesis, then her accuracy would be equivalent to that of the recognizer. Below we show the results for each corpus, averaged across the subjects:

	Switchboard	Broadcast News	Wall Street Journal
Recognizer	43.9%	27.2%	13.2%
Oracle	32.7%	22.6%	7.9%
Human Selection	42.0%	25.9%	10.1%
Human Edit	41.0%	25.2%	9.2%

Table 1 Word Error Rate: Recognizer, Oracle and Human

In the following table, we show the results as a function of what percentage of the difference between recognizer and oracle the humans are able to attain. In other words, when the human is not restricted to the 10-best list, he is able to advance 75.5% of the way between recognizer and oracle word error rate on the Wall Street Journal.

	Switchboard	Broadcast News	Wall Street Journal
Human Selection	17.0%	28.3%	58.5%
Human Edit	25.9%	43.5%	75.5%

Table 2 Human Gain Relative to Recognizer and Oracle

There are a number of interesting things to note about these results. First, they are quite encouraging, in that people are able to improve the output on all corpora. As the accuracy of the recognizer improves, the relative human improvement increases. While people can attain over three-quarters of the possible word error rate reduction over the recognizer on Wall Street Journal, they are only able to attain 25.9% of the possible reduction in Switchboard. This is probably attributable to two causes. The more varied the language is in the corpus, the harder it is for a person to predict what was said. Also, the higher the recognizer word error rate, the less reliable the contextual cues will be which the human uses to choose a lower error rate string. In Switchboard, over 40% of the words in the highest ranked hypothesis are wrong. Therefore, the human is basing her judgement on much less reliable contexts in Switchboard than in the much lower word error rate Wall Street Journal, resulting in less net improvement.

For all three corpora, allowing the person to edit the output, as opposed to being limited to pick one of the ten highest ranked hypotheses, resulted in significant gains: over 50% for Switchboard and Broadcast News, and 30% for Wall Street Journal. This indicates that within the paradigm of n-best list postprocessing, one should strongly consider methods for editing, rather than simply choosing.

In examining the relative gain over the recognizer the human was able to achieve as a function of sentence length, for the three different corpora, we observed that the general trend is that the longer the sentence is, the greater the net gain is. This is because a longer sentence provides more cues, both syntactic and semantic, that can be used in choosing the highest quality word sequence. We also observed that, other than the case of very low oracle error rate, the more difficult the task is the lower the net human gain. So both across corpora and corpus-internal, we find this relationship between quality of recognizer output and ability of a human to improve upon recognizer output.

3 Usefulness of Linguistic Information

In discussions with the participants after they ran the experiment, it was determined that all participants essentially used the same strategy. When all hypotheses appeared to be equally bad, the highest-ranking hypothesis was chosen. This is a conservative strategy that will ensure that the person does no worse than the recognizer on these difficult cases. In other cases, people tried to use linguistic knowledge to pick a hypothesis they felt was better than the highest ranked hypothesis.

In Figure 2, we show the distribution of proficiencies that were used by the subjects. We show for each of the three corpora, the percentage of 10-best instances for which the person used each type of knowledge (along with the ranking of these percentages), as well as the net gain over the recognizer accuracy that people were able to achieve by using this information source. For all three corpora, the most common (and most useful) proficiency was that of closed class word choice, for example confusing the words *in* and *and*, or confusing *than* and *that*. It is encouraging that although world knowledge was used frequently, there were many linguistic proficiencies that the person used as well. If only world knowledge accounted for the person's ability to improve upon the recognizer's output, then we might be faced with an AI-complete problem: speech recognizer improvements are possible, but we would have to essentially *solve AI* before the benefit could be realized.

One might conclude that although people were able to make significant improvements over the recognizer, we may still have to *solve linguistics* before these improvements could actually be realized by any actual computer system. However, we are encouraged that algorithms could be created that can do quite well at mimicking a number of proficiencies that contributed to the human's performance improvement. For instance, determiner choice was a factor in roughly 25% of the examples for the Wall Street Journal. There already exist algorithms for choosing the proper determiner with fairly high accuracy (Knight(1994)). Many of the cases involved confusion between a relatively small set of choices: closed class word choice, determiner choice, and preposition choice. Methods already exist for choosing the proper word from a fixed set of possibilities based upon the context in which the word appears (e.g. Golding(1996)).

Conclusion

In this paper, we have shown that humans, by postprocessing speech recognizer output, can make significant improvements in accuracy over the recognizer. The improvements increase with the recognizer's accuracy, both within a particular corpus and across corpora. This demonstrates that there is still a great deal to gain without changing the recognizer's internal models, and simply operating on the recognizer's output. This is encouraging news, as it is typically a much simpler matter to do postprocessing than to attempt to integrate a knowledge source into the recognizer itself.

We have presented a description of the proficiencies people used to make these improvements and how much each contributed to the person's success in improving over the recognizer accuracy. Many of the gains involved linguistic proficiencies that appear to be solvable (to a degree) using methods that have been recently developed in natural language processing. We hope that by honing in on the specific high-yield proficiencies that are amenable to being solved using current technology, we will finally advance beyond n-grams.

There are four primary foci of future work. First, we want to expand our study to include more people. Second, now that we have some picture as to the proficiencies used, we would like to do a more refined study at a lower level of granularity by expanding the repertoire of proficiencies the person can choose from in describing her decision process. Third, we want to move from *what* to *how*: we now have some idea what proficiencies were used and we would next like to establish to the extent we can how the human used them. Finally, eventually we can only prove the validity of our claims by actually using what we have learned to improve speech recognition, which is our ultimate goal.

References

Brill E, Harris D, Lowe S, Luo X, Rao P, Ristad E and Roukos S. (1996). *A hidden tag model for language.* In "Research Notes", Center for Language and speech processing. The Johns Hopkins University. Chapter 2.

Chelba C, Eagle D, Jelinek F, Jimenez V, Khudanpur S, Mangu L, Printz H, Ristad E, Rosenfeld R, Stolcke A and Wu D. (1997) *Structure and Performance of a Dependency Language Model.* In Eurospeech '97. Rhodes, Greece.

Chomsky N. (1956*) Three models for the description of language.* IRE Trans. On Inform. Theory. IT-2, 113-124.

Della Pietra S, Della Pietra V, Gillett J, Lafferty J, Printz H and Ures L. (1994) *Inference and Estimation of a Long-Range Trigram Model.* In Proceedings of the Second International

Colloquium on Grammatical Inference. Alicante, Spain.

Fong E and Wu D. (1995) *Learning restricted probabilistic link grammars.* IJCAI Workshop on New Approaches to Learning for Natural Language Processing, Montreal.

Golding A and Roth D. (1996) *Applying Winnow to Context-Sensitive Spelling Correction.* In Proceedings of ICML '96.

Jelinek F, Lafferty J.D. and Mercer R.L. (1992) *Basic Methods of Probabilistic Context-Free Grammars.* In "Speech Recognition and Understanding. Recent Advances, Trends, and Applications", Volume F75, 345-360. Berlin:Springer Verlag.

Jurafsky D., Wooters C, Segal J, Stolcke A, Fosler E, Tajchman G and Morgan N. (1995) *Using a stochastic context-free grammar as a language model for speech recognition.* In ICASSP '95.

Knight K and Chandler I. (1994). *Automated Postediting of Documents. Proceedings,* Twelfth National Conference on Artificial Intelligence.

Seymore K. and Rosenfeld R. (1997) *Using Story Topics for Language Model Adaptation.* In Eurospeech '97. Rhodes, Greece.

Stolcke A. (1997) *Linguistic Knowledge and Empirical Methods in Speech Recognition.* In AI Magazine, Volume 18, 25-31, No.4.

(1) people	consider	what	they	want	but	we	won't	comment	he	said
(2) people	to say	what	they	want	but	we	won't	comment	he	said
(3) people	can say	what	they	want	but	we	won't	comment	he	said
(4) people	consider	what	they	want	them	we	won't	comment	he	said
(5) people	to say	what	they	want	them	we	won't	comment	he	said

Figure 1 A sample 5-best list from the WSJ corpus. The third hypothesis is the correct one.

	Switchboard		Broadcast News		Wall Street Journal	
	% of time clicked	Absolute WER reduction using this	% of time clicked	Absolute WER reduction using this	% of time clicked	Absolute WER reduction using this
Argument Structure	1.3 (14)	0.18 (10)	2.0 (12)	0.10 (11)	5.3 (12)	0.40 (8)
Closed Class Word Choice	25.7 (1)	1.62 (1)	40.2 (1)	1.14 (1)	46.4 (1)	2.40 (1)
Complete Sent. Vs. Not	16.5 (2)	1.03 (2)	11.0 (6)	0.32 (8)	29.1 (2)	1.52 (2)
Determiner Choice	1.7 (12)	0.06 (13)	17.6 (3)	0.41 (5)	24.8 (3)	0.93 (5)
Idioms/Common Phrases	3.5 (6)	0.19 (9)	6.6 (8)	0.35 (6)	8.6 (8)	0.57 (7)
Modal Structure	2.6 (8)	0.13 (11)	3.0 (11)	0.09 (12)	2.3 (15)	0.04 (14)
Number Agreement	4.4 (5)	0.32 (8)	3.7 (10)	0.22 (9)	4.0 (14)	0.08 (13)
Open Class Word Choice	8.3 (3)	0.71 (3)	19.3 (2)	0.60 (2)	9.6 (7)	0.40 (8)
Parallel Structure	0.9 (15)	0.39 (6)	0.7 (15)	0.04 (15)	5.6 (10)	0.25 (11)
Part of Speech Confusion	2.2 (9)	0.06 (13)	2.0 (12)	0.07 (13)	7.6 (9)	0.04 (15)
Pred-Argument/Semantic Agreement	2.2 (9)	0.13 (11)	2.0 (12)	0.06 (14)	5.6 (10)	0.34 (10)
Preposition Choice	3.5 (6)	0.58 (5)	17.3 (4)	0.44 (4)	15.9 (5)	0.82 (6)
Tense Agreement	1.7 (12)	0.06 (13)	4.0 (9)	0.16 (10)	5.3 (12)	0.13 (12)
Topic	2.2 (9)	0.39 (6)	9.3 (7)	0.34 (7)	15.2 (6)	1.03 (4)
World Knowledge	6.1 (4)	0.65 (4)	12.3 (5)	0.57 (3)	19.5 (4)	1.35 (3)

Figure 2 Analysis of Proficiencies Used and their Effectiveness