Language Modeling for Good Generation

Kevin Knight
Information Sciences Institute
University of Southern California

1 Definition

This abstract refers to a technical sense of "language model"—for any English string x, the model assigns a probability P(x). We hope that the probabilities for "good" English strings are higher than those for "bad" English strings.

2 Background

At ISI, we have been using language models for some unusual tasks.

- In 1995, we started using language modeling for robust, large-scale natural language generation (NLG) of sentences from semantic input. We found it easy to write a grammar that over-generates billions of alternative sentences, but difficult to write rules that prefer one sentence over another. A language model trained on a large English corpus, however, captures many of these preferences for us. For example, "you may be required to eat chicken" is preferred over "possibly, it is necessary for you to eat chicken." In the last few years, a number of researchers have begun exploring this approach.

- Since 1997, we have also used language models for transliterating names and technical terms. Japanese and Arabic texts contain many such terms, which almost never appear in bilingual dictionaries. For example, "anji-ranaito" in Japanese must be turned into "Angela Knight" in English. We have an automatically trained "phonetic transfer" system that proposes many millions of potential English translations, which a language model must sift through. The language model prefers "Angela Knight," for example, over "Andy Law Nite" (which may be an even better phonetic fit, but seems like strange English).

- In 1999, we briefly investigated language models for deciphering writing systems. When archeologists want to decipher a script, they first aim to "make the text speak," a process similar to text-to-speech conversion, but done in the absence of any pronunciation dictionary, pronunciation rules, or even living informants. If we have a theory about the language that might be behind the writing (i.e., a language model), this drives the unsupervised learning of character-to-sound correspondences.

- We have also been using language models for sentence translation, following the pioneering work by IBM. Here, the language models responsible for selecting how to translate foreign words and phrases, and how to assemble them into a fluent English sentence.

- Most recently, we have used language models for summarization. In initial experiments, we have worked on sentence compression—which words can be dropped from a sentence so that the result preserves the important material but remains grammatical? Manually compressed sentences provide training data for determining types of important material, while the language model steers the system toward fluent output. In general, these conflict. For example, determiners are not semantically important, so we would like to drop them. Sometimes the result is still good English, but other times it is not—the language model must arbit.

3 Relevance to Information Retrieval

3.1 Transliteration

We investigated transliteration of names and technical terms with an eye toward improving automatic translation. However, it seems that this work could also have good application in cross-linguistic information retrieval (CLIR). An English name can be phonetically transferred into many Arabic, Japanese, or Chinese strings, and these can be matched against foreign-language documents.

The reverse process is also useful. For example, once we obtain a transliteration into Arabic, we can transliterate it again backwards into English. This results in many variations of the name or term, for example, "Yaser" or "Yasser" or "Yasir" or "Yassir." This is a type of query expansion that can bring in more English documents in monolingual IR.

3.2 Generation and Summarization

IR systems must understand a query, retrieve relevant information, and present the results. In the third stage, good generation is important. Retrieved information may consist of a long document, multiple documents of the same topic, etc., and we would like present the most important material in a clear and coherent manner.

Our previous models of generation and sentence compression indicate that it will be possible to produce many potential texts for the user—but that only a few of these will be grammatical and coherent. A language model must ensure that good texts are preferred.

The nice thing about the use of language models in the "noisy-channel" framework (in which we have implemented the applications described above) is that this evaluation of grammaticality and text coherence is largely independent of the overall task. They therefore attack a separable, well-defined scientific question. Sentence-level language models address the question: What makes one English sentence better than other? If we have a good answer to this, we can use it in practice to select one word over another in a given context, to select one word order over another, etc. To date, smoothed word n-gram models have proven very useful. While they leave much to be desired, they are difficult to beat.

At present, there are no satisfactory text-level language models that address this scientific question: What makes one English text more coherent than other? Clearly, scrambling the sentences in a paragraph reduces coherence. Likewise, gluing together ten sentences from ten different newspaper articles results in an incoherent text. But exactly why? The large literature on text coherence is somewhat similar to large literature on English syntax—it has not yet produced a workable, practical language model. We need to do some basic modeling here akin to what has been done with n-grams in sentence-level language models.

A model of coherence will be interesting in its own right, and there will be many applications. An important application for IR is multidocument summarization. Here, we can turn a large set of input documents into many different short summaries, and these can be ranked for coherence. If we have already synthesized a document of questionable coherence, then we can consider making changes that improve its coherence. Because the model is quantitative, these comparisons and changes are based on numerical scores, and many sources of information can be brought to bear simultaneously. Models of coherence could also be used for essay grading and other tasks.

We are optimistic about this possibility largely because vast amounts of training data exist. Every time a person writes a paragraph or document, we have a positive example of a coherent text. (Of course, some texts are more coherent than others, just as some sentences used to train word-level language models are more grammatical than others.) It is easy to construct negative examples as well. We also see a good situation with respect to test data. Models should assign high probability to previously unseen texts, and models should be able to repair texts that have been intentionally damaged. This sort of objective evaluation could lead to much speedier progress in this field.

4 Acknowledgments

Section 2 is based on joint work with Irene Langkilde, Jon Graehl, Kenji Yamada, and Daniel Marcu. Section 3 is based on joint work with Daniel Marcu.