Corpus-Based Methods and Hand-Built Methods

Richard Sproat

AT&T Labs — Research
rws@research.att.com

ABSTRACT

Recent success of statistical corpus-based methods in a variety of areas of speech and language processing has led to the widespread view that traditional hand-built “rule-based” approaches are moribund. This is a misconception. As I shall argue in this talk, it is unlikely that rule-based approaches will ever be eliminated. Two examples are given to support this conclusion; one where the linguistic facts, though highly complex are basically quite regular; and another where the linguistic fact is exceedingly simple (hence hardly worth the effort of inferring from data), but where adding in this information can improve the output of a statistical model.

1. INTRODUCTION

Statistical methods have been applied to a variety of problems in speech and language processing, and the overall consensus is that such methods are successful — so successful in fact, that they have all but supplanted more traditional “rule-based” approaches in many areas.

One such area is so-called “letter-to-sound” conversion where the problem is to convert a word written in its normal orthographic representation into a phonetic representation for that word. For English, in particular, this is a hard problem due to the complexity of the English spelling system: this is true even for ordinary English words, in part because English spelling tends to be conservative, not reflecting present-day pronunciation particularly well; but it is especially true for names, where the tendency is to spell names according to their language of origin no matter what their adapted pronunciation in English may be. Traditional approaches to the problem, such as [2], combine dictionaries with hand-crafted letter-to-sound rules though other techniques (e.g. analogical reasoning) may be used. While such approaches have achieved good coverage (an error rate of 5% or less by word is typical), it is recognized that such techniques are labor-intensive and can also be hard to maintain. It is natural therefore to seek alternative approaches that are less costly, and as a consequence of this there has been a veritable cottage industry of statistical approaches to the problem of letter-to-sound conversion starting with [10] and including more recently [1, 6, 3, 7]. Such techniques have also shown some success (though the typical error rate of 5% by phoneme is somewhat troubling), and as a consequence of this there seems to be a widespread belief that rule-based or hand-built methods are moribund.

Nonetheless, despite the importance of statistical corpus-based methods, I believe it is a misconception that they have supplanted — or will ever completely supplant — more traditional linguistic methods. The reason for believing this is that in addition to statistical tendencies, language does contain facts which are quite regular and deterministic, and which are more or less easily stated in terms of linguistically-informed hand-built rules or principles. I briefly offer two rather different examples in this paper. In the first case (Section 2.), the rules or principles involved are quite complex, and in this case it would even be difficult to see how the system could be inferred from uncooked data. Yet at the same time the phenomenon is at its core quite regular. In contrast there are regular phenomena that are quite simple to state by hand (so that it would hardly be worth the effort of inferring them), but which nonetheless can contribute to a measurable improvement in a system that is corpus-trained; this I exemplify in Section 3..

2. COMPLEX REGULAR PHENOMENA

Some linguistic phenomena are sufficiently complex that it is hard to imagine having enough of the right kind of data that would allow current inference methods to infer a correct model. An example of such a phenomenon is the reading of percentage and other numeral expressions in Russian, discussed more fully in [11].

In English the percentage symbol ‘%’, when denoting a percentage, is always read as percent. In contrast, in Russian selecting the correct form depends on a number of contextual factors, the combination of which results in a quite complex, though completely regular system. The first question that needs to be asked is whether or not the number-percent string modifies a following noun since Russian in general disallows noun-noun modification: in constructions equivalent to noun-noun compounds in English, the noun must be converted into an adjective. This general constraint applies equally to procent ‘percent’, so that the correct rendition of 20% skidka (20% скидка) ‘twenty percent discount’ is dvadcati-procentnaja skidka (двадцатипроцентная скидка) (twenty gen.-Percent+adj[nom,sg, fem] discount[nom,sg,fem]). Not only does procent have to be in the adjectival form, but as with any Russian adjective it must also agree in number, case and gender with the following noun. Observe also that the word for ‘twenty’ must occur in the genitive case: with some well-defined exceptions, numbers that modify adjectives in Russian must occur in the genitive case.

If the percentage expression does not modify a following noun, then the nominal form procent is used. The exact form used depends in turn on other factors: if the expression is in an oblique case (e.g., is governed by a preposition or verb that requires other than the nominative or accusative case), then both the number and procent show up in that case, with procent being in the singular if the number ends in one (including compound numbers like twenty one), and the plural otherwise: s odnim procentom (с одним процентом) (with one[instr, sg, nom] percent[instr, sg]) ‘with one percent’; s pятью procentami (с пятью процентами) (with five[instr, sg] percent[instr, sg]) ‘with five percent’. If the case is non-oblique, then the form of procent depends upon what the number ends with: with numbers ending in one, procent occurs in
the nominative singular. For numbers ending in so-called paucal numbers — two, three, four and their compounds — the genitive singular procenta is used. After all other numbers one finds the genitive plural procentov. So we have odin procent (один процент) (one percent), dva procenta (два процента) (two percent), and pyat’ procentov (пять процентов) (five percent). As with the adjectival forms, there is nothing peculiar about the behavior of the noun procent: all nouns exhibit similar behavior in combination with numbers [4]. What is complex about this system from the point of view of speech technology is that the written form — e.g. 20% skidka — gives no overt indication of what linguistic form it corresponds to, and the latter must be inferred using rules that model the linguistic facts described above. These facts are further summarized in Figure 1.

The Russian system is evidently very complex, yet at the same time it is very regular. Indeed it turns out that it is possible to implement it using a relatively small set of finite-state “local grammar” constraints, as described in [11]. It is not clear it would be worth the effort to try to prepare enough Russian data to train a statistical method to infer these constraints. To be sure, statistical methods could be used to address some aspects of this problem. For instance, deciding whether or not a percentage expression is likely to be a modifier of another noun is akin to the problem of part-of-speech disambiguation (effectively: is the percentage expression to be tagged as an adjective or a noun?) and just as statistical methods have proved very effective in part-of-speech tagging, one would expect them to also prove effective here. Similarly, one might expect a statistical model of the context to be beneficial in deciding whether a nominal percentage expression should be reflected in an oblique case or not. But once those points are decided, the rest of the computation, though complex, is quite regular, and is arguably best done using linguistically-informed rules.

3. COMBINING RULES WITH STATISTICAL METHODS

Some linguistic generalizations may, in contrast, be very simple to state — so simple that it would hardly be worth the effort of trying to infer them statistically from data, but which nonetheless can contribute significantly to improving the output of an otherwise statistical model. I give a simple example of this in this section.

First of all we describe a statistical method (one of many) for training English grapheme/phoneme correspondences. We started with a dictionary of English words and their pronunciations. The dictionary was derived from an earlier edition of the Collins English dictionary with hand-corrected pronunciations (52,000 words); a dictionary of 50,000 surnames, also with hand-corrected pronunciations; 317,000 words from the XTAG dictionary (from the Xtag project http://www.cis.upenn.edu/~xtag/) with pronunciations automatically generated using the pronunciation module described in [2]; and 6,300 additional words from a corpus of text read by a speaker for a unit-selection-based text-to-speech system, again with automatically-derived pronunciations. The union of these yielded a total of 335,000 distinct words. The pronunciation information in this dictionary included stress information (primary and secondary stress) and syllable boundary information. 268,000 words were selected as training data.

For this training data, we computed an alignment of graphemes with corresponding phonemes using an automatic alignment procedure. We then constructed a 6-gram language model with Katz backoff [5] over the pairs of symbols. Using the AT&T Weighted finite-state transducer (WFST) toolkit [9, 8], such language models can be represented as weighted finite-state acceptors (the particular implementation of this being due to Don Hindle). In this case the acceptor is over pairs of symbols, but this can easily be converted into a WFST that maps between graphemes and possible phoneme sequences. Pronunciation prediction consists of composing this WFST — call it T — with the word, represented as an acceptor W, projecting this onto the righthand (pronunciation-side) dimension, and selecting the cheapest path in the lattice. Formally:

$$\text{BestPath}(\pi_2(W \circ T))$$

We tested the WFST on a held-out randomly selected 1000-word set. Ignoring stress and syllable marks (as is typical) yielded a phoneme error rate of 5%, and a word error rate of 19.7% (comparable to the best reported statistical methods for English). Counting stress and syllable boundaries yields a phoneme error rate of 8% and a word error rate of 31%.

Now, many stress errors involve not predicting any primary stress, or predicting more than one; this is expected since a 6-gram model is not wide enough to “see” the entire stress context. But it is straightforward to impose an additional constraint — implementable as an FSA S — requiring exactly one primary stress per word: note that it is virtually a corollary of the definition of primary stress that there be exactly one and only one primary stressed syllable in a word. This FSA is simply intersected with the output of the composition before the cheapest path is computed:

$$\text{BestPath}(\pi_2(W \circ T) \cap S)$$

Doing this results in a 7% phoneme error rate and a 23% word error rate — a 26% reduction in word error rate. The filter even affords a slight error rate reduction when one ignores stress in the evaluation (e.g. a 19.1% versus 19.7% word error rate). These results are summarized in Table 2.

To be sure this method of combining linguistically motivated information with a statistical model suffers from some drawbacks. In particular, in filtering out output, we are also removing probability mass, so that although the analyses preserved by the filter are still ranked correctly in terms of relative probabilities, the probabilities are underestimated. One may, for example, try to solve this problem by renormalizing, in some fashion, the probability mass, so that although the analyses preserved by the filter are still ranked correctly in terms of relative probabilities, the probabilities are underestimated. One may, for example, try to solve this problem by renormalizing, in some fashion, the probability mass, so that although the analyses preserved by the filter are still ranked correctly in terms of relative probabilities, the probabilities are underestimated. One may, for example, try to solve this problem by renormalizing, in some fashion, the probability mass, so that although the analyses preserved by the filter are still ranked correctly in terms of relative probabilities, the probabilities are underestimated.
if the percentage phrase is modifying a noun then use the "procentn-" adjectival form, agreeing in case, number and gender with the noun and put the number in the genitive case (for most numbers)

else use the "procent" nominal form

if the percent phrase is in an oblique case, then inflect the number and "procent" for that case, and make "procent" singular if the number ends in "one", plural otherwise.

else if the number ends in "one" inflect "procent" in the nominative singular

else if the number ends in a paucal number inflect "procent" in the genitive singular

else inflect "procent" in the genitive plural

Table 1: How to decide how to say a percentage term in Russian.

<table>
<thead>
<tr>
<th>Condition</th>
<th>PER</th>
<th>WER</th>
</tr>
</thead>
<tbody>
<tr>
<td>NSS</td>
<td>0.05</td>
<td>0.197</td>
</tr>
<tr>
<td>SS</td>
<td>0.08</td>
<td>0.31</td>
</tr>
<tr>
<td>filt + NSS</td>
<td>0.05</td>
<td>0.191</td>
</tr>
<tr>
<td>filt + SS</td>
<td>0.07</td>
<td>0.23</td>
</tr>
</tbody>
</table>

Table 2: Summary of results of English word-pronunciation experiment. Conditions: NSS, not counting stress and syllable boundaries in evaluation; SS, counting stress and syllable boundary; filt, using FSA stress filter on output. PER = phoneme error rate; WER = word error rate.

4. DISCUSSION

We have presented two brief case studies that attempt to show that, despite the strong focus in recent years on statistical corpus-based methods in speech and language processing, more traditional linguistic approaches still have important things to contribute. The arguments outlined in this paper should not be interpreted as an attack on corpus-based statistical methods. On the contrary, the correct conclusion to draw is that speech and language systems will continue to be a marriage of both statistical and hand-built approaches. A crucial key to this marriage is a computational framework — of which the AT&T WFST toolkit is one instance — that can naturally implement and combine models of both kinds.

5. REFERENCES