ON THE FEASIBILITY OF AUTOMATIC PUNCTUATION OF TRANSCRIBED SPEECH WITHOUT PROSODY OR PARSING

Mary O'Kane, P.E.Kenne, Hamish Pearcy
The University of Adelaide
and
Tim Morgan, Gail Ransom and Kathryn Devoy
University of Canberra

ABSTRACT - This paper describes an investigation of the effectiveness of statistical methods for automatic punctuation of transcribed speech. Most work carried out on automatic punctuation is based on prosodic or syntactic analysis. Here, however, we decided to investigate the strengths and weaknesses of automatic punctuation based solely on the more simply-calculated probabilities of the collocation of certain words or groups of words with different punctuation marks augmented by some simple heuristic rules. Using the techniques described in this paper, just over half the correct positions of punctuation marks can be found and about 42% of all the expected punctuation results are correctly marked at the cost of 7% of the total number of expected marks being incorrect insertions. It has been found that when deriving the statistical training data for the punctuator, it is important to use text of the same style as that which is to be punctuated automatically.

INTRODUCTION

The phenomenon of punctuation is generally regarded as either a representation of the prosody of the spoken word or as an integral part of the syntax of the language. It follows from this that the way to punctuate automatically the text generated by a speech recogniser should be to have access either to the prosody of the spoken words, or to some syntactic representation of the output text, or both. While humans clearly rely on prosody (intonation, stress, duration, etc.) to punctuate heard speech, automatic extraction of prosody is acknowledged to be a difficult problem that is currently being investigated in several large projects in Australia (Eggins, Vonwiller, Matthiesson and Sefton, 1991) and overseas (Goodine, Seneff, Hirschmann and Phillips, 1991). Automatic parsing of spoken language is also a difficult problem. For these reasons we decided to investigate how effective it was to use easily-derived, statistical, word-level knowledge for punctuating spoken language passages automatically. Such knowledge does not rely on issues that are as difficult to extract as prosody-based or parsing-based measures. The real challenge of this project is to see how effective non-prosody-based , non-parsing-based measures can be in achieving automatically correct and comprehensive punctuation of machine-transcribed text.

The primary data used in this study were court transcripts. This transcribed speech belongs admittedly to a specialised genre, but it is of particular interest as it is a (close to) verbatim transcription of the speech of participants in a dialogue (apart from certain stylistic conventions used in transcription e.g. the omission of indications of stuttering). Automatic punctuation is of interest as a problem because as machine recognition of human speech is used more often to transcribe speech to text, there is a need then for the resulting string of words to be turned into properly-punctuated, natural-language sentences.

The initial attempt at automatic punctuation of raw transcript was based on the probabilities of the collocation of certain words or groups of words with different punctuation marks derived from a study of large amounts of existing punctuated transcript. These probabilities were then used to insert punctuation into an unpunctuated text. The correctness or otherwise was measured using four indices: <right place, any mark>, <right place, right mark>, <marks missed> and <incorrect insertions>. The success of the punctuation using simple collocation probabilities was, not surprisingly, found to be dependent on the amount of training transcript used to derive the collocation probabilities. Thus for probabilities derived from analysis of a 200,000 word court transcript excerpt, and tested on transcript from another court, the right punctuation mark is inserted only 21% of the time. This figure rises to 35% for probabilities derived from a 1,400,000 word excerpt. This works similarly when the punctuator is tested on a widely varying range of court transcripts. However the results are only about half as good when the punctuator is tested on a totally different style of English (as is to be found in a novel).

After describing the training and testing of the automatic punctuation system, we give an example of output from the punctuator and discuss points raised by it, particularly what exactly is expected of such a system.

THE NATURE OF THE PROBLEM

As mentioned above, the primary training and test data used in this work were transcribed court proceedings. Transcribed speech illustrates the well-known fact that speakers do not always use complete well-formed sentences and, in dialogue, rely heavily on anaphorically-derived sentence abbreviations. Punctuation of such text is generally more challenging than punctuation of text from, say, a scientific paper. Consider the following court extract with punctuation and capitalisation removed (note that speech from different speakers can be distinguished because they are speaking into different microphones and being recorded on different channels).

> LAWYER 1: and you'd asked for it and it wasn't forthcoming is that correct
>
> THE WITNESS: not in physical terms mate there would have been advice given you know I mean well you can delete this or include such-and-such and just rephrase certain clauses within the award.
>
> LAWYER 1: but you're being critical of the victorian federal union here aren't you
>
> THE WITNESS: well certainly as it appears on this bulletin yes i am
>
> LAWYER 1: and that was the fact wasn't it that you were critical of the federal victorian union
>
> LAWYER 2: well the document speaks for itself
>
> LAWYER 1: well i'm entitled to cross-examine your honour
>
> HIS HONOUR: he's entitled to ask that question mr X
>
> THE WITNESS: i'm being critical of the federal sorry the victorian/federal union in this document
>
> LAWYER 1: and you were being critical of them not for propaganda purposes but because as a fact you were critical of what they had done for you
>
> THE WITNESS: yes

The significance of the work described here is that we aim to produce a system that would take input text such as shown in the example above and would produce as output an approximation to the text shown below which is the court transcript as transcribed and punctuated by Auscript, the Australian Court Reporting Service. It should be noted that one can debate what is meant by 'correct' punctuation. It is a useful exercise to try to punctuate the unpunctuated text above and then to compare one's results with the 'official' version given below. We will return to this issue of correctness below.

> LAWYER 1: And you'd asked for it, and it wasn't forthcoming; is that correct?
>
> THE WITNESS: Not in physical terms, mate; there would have been advice given, you know, I mean, well you can delete this, or include such-and-such, and just rephrase certain clauses within the award.
>
> LAWYER 1: But you're being critical of the Victorian federal union here, aren't you?
>
> THE WITNESS: Well, certainly as it appears on this bulletin, yes, I am.
>
> LAWYER 1: And that was the fact, wasn't it, that you were critical of the federal Victorian union.

LAWYER 2: Well, the document speaks for itself.

LAWYER 1: Well, I'm entitled to cross-examine, your Honour.

HIS HONOUR: He's entitled to ask that question, Mr X.

THE WITNESS: I'm being critical of the federal - sorry, the Victorian/Federal union in this document.

LAWYER 1: And you were being critical of them not for propaganda purposes, but because, as a fact you were critical of what they had done for you?

THE WITNESS: Yes.

THE PUNCTUATOR

The punctuator reads in unpunctuated text and makes a best guess at the punctuation marks which should occur between the words by examining statistical data from human-punctuated text.

The program can take input from 4 different files, any of them may be omitted through command-line options of the program:

- pos.dat is position data. This file contains probabilities for each punctuation symbol in a given position. Position is currently interpreted as the number of words since the last time the symbol occurred, and the probabilities are cumulative, i.e. the probability value at a position is the probability that the symbol has occurred somewhere between that position and the last position the symbol was inserted;

- pair.dat is word pair data. This is a file created by the pair spot program, and contains punctuation information for known word pairs. Each line is of the format <word1 word2 I punct I count I total> where word1 and word2 are the words that make up the pair, punct is the punctuation *between* the two words, count is the number of times this punctuation was discovered in the training data with this word pair, and total is the total number of times the word pair occurred in the training data. As this data is read in, a probability for each punctuation symbol in each word pair is calculated by dividing count by total.;

- bef.dat is data for punctuation *before* single words. It is created by the pair spot program and has the same format as the word pair data, except that there is only one word;

- aft.dat is data for punctuation *after* single words. It is created by the pair spot and has the same format as the before data.

In its second pass the punctuator uses a set of simple heuristic rules such as checking that sentences beginning with an interrogative end with a question mark.

The unpunctuated text is fed to the punctuator as shown in Figure 1. The punctuator returns the best guess for a punctuation symbol, or NULL if none can be determined. It attempts to calculate a probability score for each punctuation symbol, and choses the one with the highest score. As well as producing punctuated text, the punctuator also collects a series of performance statistics (when it has access to the human-produced 'correct' punctuated text):

- <right place, any mark>, a count of all the punctuation that was inserted by the punctuator in a place where punctuation should have occurred, regardless of what the actual punctuation mark was:

- <right place, right mark>, a count of all correctly-inserted punctuation;

- <place missed>, a count of the punctuation places in the original Auscript file that were missed by he punctuator;

- <wrong place>, a count of the number of times punctuation was inserted where no punctuation should have been.

All of the above are often expressed as a percentage of the total number of punctuation marks in the Auscript file.

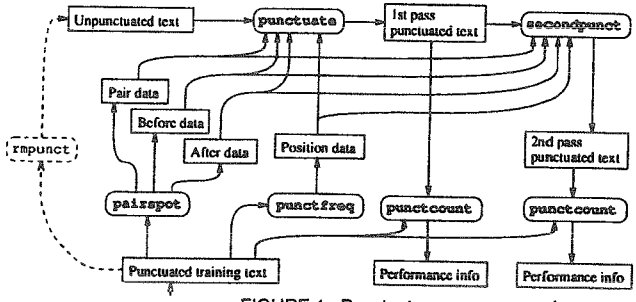

FIGURE 1: Punctuator program overview

Using these performance measures we can examine the relative effectiveness of the various different types of statistical input data that are (optionally) used by the punctuator. In Figure 2 we show the results for the punctuator trained in various ways on one court case and then tested on another case. A surprising result is that including position data actually seems to degrade performance.

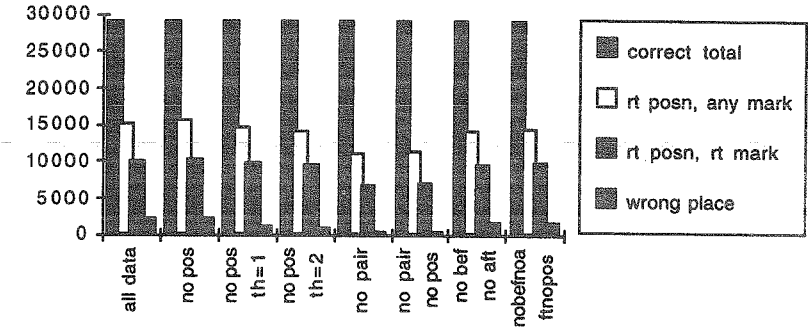

FIGURE 2: Punctuator results from training using different options available in the Punctuator program.

EFFECT OF TRAINING SET SIZE

In order to try to establish the minimum amount of training data needed for the punctuator to achieve optimal performance, we examined how the performance statistics for the punctuator improved as the amount of training data was increased. The results of doing this by training on successively larger and larger amounts of data from one file and testing on data from another court transcript file can be seen in Figure 3. As can be seen from the diagram, after about one million words in the training file, the increase in punctuator performance in terms of increased right place and right mark and decreased place missed and wrong place is very slow for the same increase in training set size. This is not surprising because while a very small number of words and word pairs account for the bulk (about 80%) of the text, a very large number of words account for the remainder (O'Kane, 1993).

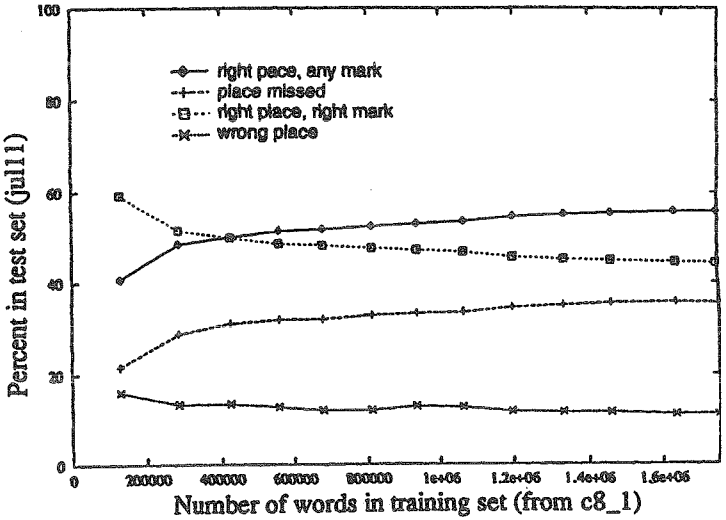

FIGURE 3: Punctuator performance improvement as training set size is increased.

EFFECT OF TRAINING DATA TYPE

In investigating punctuator performance increase with training set size, we noticed that the increase in punctuator performance is faster, the more alike are the training and test sets and *vice versa*. Thus if the punctuator is trained successively on the first days of a trial and tested on the last days of the same trial the performance improvement is faster than if the testing is done on another trial. This makes good sense as the vocabulary of any particular trial will be somewhat different to the vocabulary of an another trial.

An obvious further issue to investigate in this regard is what happens if the training and testing for the punctuator are done on very different styles of text. We examined this by training the punctuator on the first day's proceedings of one court case and tested it on both one day of proceedings from another trial and an extract form Dracula (Stoker, 1897). The results were as shown in Table 1. Clearly the very different text style (prose with well-formed sentences as opposed to dialogue with many short and interrupted sentences) and the very different vocabularies mean that a statistical punctuator such as this is critically dependent on appropriate training material.

	Right Place, Any Mark	Right Place, Right Mark	Place Missed	Wrong Place
Court test	53.1%	33.8%	46.9%	8.1%
Dracula test	18.8%	9.9%	81.2%	12.5%

TABLE 1: Comparison of court-trained punctuator tested on another court extract and on an extract from Dracula.

WHAT IT ACTUALLY LOOKS LIKE

To provide a more immediate sense of what the final result of the punctuator is like we produce the example given above after it has been punctuated by the punctuator. Although the punctuator has an optional switch for capitalisation we have not used it here as we are concentrating on punctuation *per se* . The punctuator was trained in this case on one court case and the extract comes from a totally separate court case but within the same jurisdiction (the Federal Industrial Court). The performance

results for the case from which this excerpt was taken are <right place, any mark> 63.3%; <right place, right mark> 41.8%; <place missed> 36.7%; <wrong place> 7.0%.

LAWYER 1: and you'd asked for it and it wasn't forthcoming is that correct.

THE WITNESS: not in physical terms mate there would have been advice given you know, i mean - well, you can delete this or include such-and-such and just rephrase certain clauses within the award - - -

LAWYER 1: but you're being critical of the victorian federal union here, aren't you?

THE WITNESS: well, certainly as it appears on this bulletin yes, i am.

LAWYER 1: and that was the fact wasn't it that you were critical of the federal victorian union?

LAWYER 2: well, the document speaks for itself.

LAWYER 1: well, i'm entitled to cross-examine your honour.

HIS HONOUR: he's entitled to ask that question, mr X?

THE WITNESS i'm being critical of the federal sorry, the victorian/federal union in this document.

LAWYER 1: and you were being critical of them not for propaganda purposes, but because as a fact you were critical of what they had done for you?

THE WITNESS: yes.

Considering this example and comparing it with the 'correct' Auscript version given earlier, we notice that the punctuator has not ended the first question from Lawyer 1 with a question mark and has incorrectly put a question mark at the end of the judge's assertion in the middle of the example. On the other hand, the Auscript version has a mistake in that it does not have a question mark at the end of the third Lawyer 1 question while the punctuator gets this correctly. While there are punctuation marks missing in the output from the punctuator, it is interesting to consider if the punctuation that has been inserted is sufficient for an 'adequate' understanding of the sense of the dialogue. Also the question of correctness involves an overall sense of the dialogue, so what might be correct locally is not necessarily correct in a more global sense. Thus, the question mark inserted by the punctuator at the end of the comment by the judge is just as correct as a full stop when the sentence is considered in isolation but is not correct when the sentence is taken in the context of the complete dialogue. This example gives some indication of the limits of a statistically-based punctuator. It is unlikely that such a system would ever be able to resolve such a case without reference to prosody or semantic/dialogue analysis.

REFERENCES

Eggins S.,Vonwiller J., Matthiessen C. M. I. and Sefton P. (1991) "The description of minor clauses in information-seeking telephone dialogues", Proceedings Eurospeech'91, Genova, 1059-1062.

Goddine D., Seneff S., Hirschman L. and Phillips M. (1991) "Full integration of speech and language understanding in the MIT spoken language system", Proceedings Eurospeech'91, Genova, 845-848.

O'Kane M. J. (1993) "Listening intelligently and giving a sensible answer" in *AI'93* pp. 3-11 (World Scientific: Singapore; C. Rowles, H. Liu and N. Foo, eds.).

Stoker B. (1897 - repub. 1966) *Dracula* (Hutchinson:London).