The Effects of Speech Recognition and Punctuation on Information Extraction Performance

John Makhoul, Alex Baron, Ivan Bulyko, Long Nguyen, Lance Ramshaw, David Stallard, Richard Schwartz, Bing Xiang

BBN Technologies
Cambridge, MA 02138, USA
makhoul@bbn.com

Abstract

We report on experiments to measure the effect of speech recognition errors and automatic punctuation insertion errors on the performance of information extraction (entity and relation extraction). The outputs of several recognition systems with a range of word error rates (WER), along with punctuation insertion, were fed into a system that extracts entities and relations from the recognized text. Entity and relation value scores were measured as a function of WER and types of punctuation used. The results of the experiments showed that both entity and relation value scores degrade linearly with increasing WER, with a relative reduction in scores of about twice the WER. The information extraction modules require the inclusion of sentence boundaries, at a minimum; however, the experiments showed that the exact locations of these boundaries are not important for entity and relation extraction. In contrast, when comparing the effects of full punctuation to just automatic sentence boundary insertion, there was a loss in entity value scores of 13.5% and in relation value scores of 25%. Further, commas play a significantly greater role in entity and relation extraction than other types of punctuation.

1. Introduction

Most work on the extraction of entities and the relations among those entities assumes that the input to the extraction is running text that is punctuated and that contains few if any errors. However, for many applications, it is important to extract such information from spoken input. In a previous study, Miller *et al.* [1],[2] measured the effect of speech recognition errors on the extraction of named entities (such as names of persons, locations, and organizations) from speech and found that the F-measure degraded at a rate of 0.7 of the WER (and the entity error rate degraded approximately one-for-one with recognition errors).

In this study, we examined the effects of recognition errors and several types of punctuation on the performance of a more complicated information extraction task, that of entity and relation extraction. The entities here are more general than named entities and include co-reference resolution. For example, "George W. Bush", "the President", and any pronouns that refer to President Bush, all have to be identified with the same entity. Extracting an entity correctly requires both finding and linking the correct set of mentions of the entity. Relations, in turn, connect two entities with one another; one example would be the relation "works-for", which can hold between a person and an organization. Clearly, finding relations among entities is a significantly more difficult task than just finding the entities.

In this study, speech is first processed through one of a number of recognizers in order to simulate various word error rates. The output text string from the speech recognition is then punctuated either automatically or by inserting the punctuation from the reference text, which had been punctuated manually. The text, thus punctuated, is then processed through the information extraction system to produce a set of entities and relations, which are then scored against a manually generated reference. The result is a set of entity and relation value scores. We report here the results of this study.

Section 2 describes the corpus used for the study. Section 3 describes the systems used to produce output with different word error rates to be used as input to the information extraction system. Section 4 describes the component used to insert punctuation, and the methods used to create reference conditions for the experiments. Section 5 gives a brief description of the information extraction system used in this study and the evaluation measures used to assess the performance of the system. Section 6 presents and discusses the results of the various experiments.

2. Corpus

The corpus comprised a set of 946 articles in English that were taken from DARPA's Automatic Content Extraction (ACE) program [3], which dealt with the extraction of entities and relations from text. The articles were those used in the ACE 2002, 2003, and 2004 evaluations – all are available from the Linguistic Data Consortium (LDC). Of these articles, 578 articles were closed-caption transcriptions of broadcast news stories that were part of the TDT corpus and the remaining 368 articles were newswire articles. Speech for the broadcast news stories (15 hours) was already available, but we did not have a spoken version of the newswire articles. Those articles were read and recorded by LDC resulting in 19.2 hours of speech. The breakdown of the resulting corpus is shown in Table 1.

The reference texts corresponding to this speech corpus were derived from the original mixed-case, punctuated corpus, by rendering it single case, maintaining the original punctuation, and converting all numerals and other special symbols to their spoken versions. It is worth noting that, for both parts of the speech corpus, the reference texts were not exact transcripts of the audio. For the TDT data, the reason was that the text was

a closed-captioned version of the audio and, for the newswire data, the speakers made some errors while reading the articles. Nevertheless, we measured the recognition WER on the available reference text data since that is what the information extraction system saw.

Data	# articles	# hours
TDT	578	15.5
Newswire	368	19.2
All	946	34.7

Table 1: Configuration of the speech corpus.

3. Speech Recognition

To quantify the effects of speech recognition errors on information extraction at different WER, we used four broadcast news (BN) English systems to transcribe the audio. As shown in Table 2, the WERs ranged from 7.6% to 24.1%.

System	Training Data (hrs)	TDT WER (%)	Newswire WER (%)	Average WER (%)
I	1700	8.9	6.6	7.6
II	1700	11.7	10.0	10.8
III	140	19.2	16.5	17.7
IV	140	25.4	23.1	24.1

Table 2: Speech recognition results for the four systems used in the study.

System II is BBN's 10xRT system used in the DARPA EARS 2004 evaluation; it was trained on 1700 hours of speech data. System I was derived from System II by including the test data in the language model with appropriate weights. System III is the one BBN had at the beginning of the EARS program in 2002; it was trained with 140 hours of speech only. System IV, with the highest WER, was derived from System III by decreasing the grammar weight during decoding.

4. Punctuation Insertion

The information extraction software requires at a minimum sentence breaks (periods, question marks, or exclamation points) in order to run. It also benefits from additional punctuation marks, such as commas for certain common constructions. Accordingly, we developed software to insert punctuation into the recognition output. The software was used to insert various levels of punctuation taken from the reference texts. We also included, as one experimental condition, the use of an automatic sentence boundary detector.

To insert punctuation from the reference text, we aligned the words of the reference text to the speech recognition output. Each word in the reference was associated with any punctuation attached to it. After the reference and recognition output words were aligned, each word in the recognition output was then punctuated according to the word with which it aligned in the reference. If a reference word was deleted, its punctuation was applied to the preceding recognition output word in the alignment.

For our automatic sentence boundary detector, we used the SRI LM toolkit [5] to implement the algorithm proposed by SRI [6]. All sentence-final punctuation was converted to periods during training and a standard 3-gram language model (LM) was built as a mixture of LMs constructed from three data sources: TDT3 closed captions, Hub4 broadcast news transcripts, and Gigaword articles dated from 2000 – a total of about 250M words of training data. This language model was then used to predict locations of sentence boundaries yielding a state-of-the-art punctuation error rate of 61% when applied to current state-of-the-art recognition output. The sentence boundary error rates increased slowly but linearly as the recognition error increased, ranging from 58% on speech reference to 68% on recognition output with 24.1% WER.

5. Information Extraction

The particular information extraction system we used was BBN's SERIF system [4], which uses statistical name-finding and parsing algorithms to identify entity mentions in the text, and statistical evidence plus supplemental rules to identify relations. It has been among the leading systems in ACE evaluations. For this study, the system was trained on mono-case text.

In ACE evaluations, the Entity Value Score (EVS) is used to measure the accuracy of entity extraction; it is based on a weighted percentage of entity mentions that were correctly identified and grouped into entities.[7] Similarly, the Relation Value Score (RVS) is derived from the percentage of relations that were correctly identified. Entity mentions are identified in terms of the character offsets in the document at which the mention was judged to begin and end. For an entity mention to be judged correct, the character offset intervals of the reference and candidate entity must have at least a 30% overlap. The actual text of the entity mention is not checked because the software assumes that the input text is correct.

One problem with this scoring methodology for speech is that incidental differences in white space between the original text document and recognition output, plus recognition errors, render a character interval comparison meaningless. We solved this problem by aligning the words of the document and recognition output, and using this alignment to map the character positions of the recognition output to the character positions of the document.

A second problem was that, while the scoring procedure compares text intervals, it does not compare the actual words in those intervals. So, it is possible that a word error in the recognition output will change the entity being recognized, yet still pass the overlap condition and be scored as correct. We dealt with this problem by adding a requirement to the scoring procedure that at least 50% of the alphanumeric characters in the candidate and reference text intervals match.

6. Results and Discussion

Table 3 shows the EVS and RVS for various conditions of speech recognition WER and punctuation settings. In the *All Punct* condition, all punctuation marks that existed in the reference texts were inserted into the speech recognition output via forced alignment. In the *Period Only* condition, only reference sentence boundaries were inserted, including question marks and exclamation points (which were converted to periods). In the *Period,Comma* condition, both reference sentence boundaries and commas and semicolons were inserted (the latter being converted to commas). Finally, in the *Auto Period* condition, only periods were inserted automatically, using the method described in Section 4.

6.1. Effects of Punctuation

Sentence boundaries are very important for information extraction, simply because the software will not run without them. However, it appears that the exact placement of those sentence boundaries is not as important as one might expect, at least as far as the entity and relation value scores are concerned. That can be seen by comparing the *Period Only* and *Auto Period* results for WER=0% and WER=10.8% in Table 3. In both cases, the reduction in EVS and RVS is on the order of a few percent relative, even though the punctuation error in automatic boundary placement is on the order of 60%! The results for EVS can also be seen graphically in the bar chart in Figure 1.

The importance of commas for information extraction can be seen by comparing the *Period,Comma* results (where reference sentence boundaries and commas were inserted) with the *Period Only* results (where only reference sentence boundaries were inserted) for both WER=0% and WER=10.8%. This is illustrated in both Table 3 and the bar chart in Figure 1. By not including commas, the EVS drops around 9.5% and the RVS drops by about 15% relative, both very substantial drops. This significant effect can be explained in part by the importance of appositives, which are marked by commas, in detecting certain entities. (An example of an appositive is: "George W. Bush, the President of the United States, said this morning …") Appositives are found by rule and are very dependent on having the correct commas there.

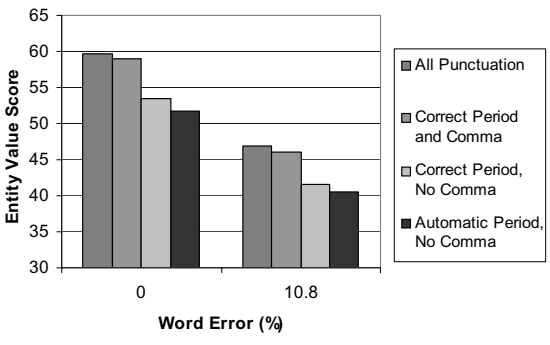

Figure 1: Effects of punctuation on entity value scores.

Inserting all punctuation marks (not just sentence boundaries and commas) does improve the IE results, but more so for relations than for entities, as can be seen at all word error rates in Table 3.

Overall, going from full correct punctuation to using only automatic boundary detection reduces EVS by 13.5% on average. The drop in RVS is more erratic, with an average of about 25%. Therefore, being able to insert correct punctuation has a significant effect on automated information extraction.

WER	Punctuation	EVS (%)	RVS(%)
0.0%	All Punct	59.7	27.3
	Period,Comma	58.9	25.9
	Period Only	53.4	21.3
	Auto Period	51.7	19.9
7.6%	All Punct	49.3	22.9
	Period,Comma	48.5	21.4
	Auto Period	42.7	17.8
10.8%	All Punct	46.9	22.0
	Period,Comma	46.0	20.5
	Period Only	41.5	17.7
	Auto Period	40.5	17.4
17.7%	All Punct	39.1	17.8
	Period,Comma	38.0	16.6
	Auto Period	33.9	14.3
24.1%	All Punct	31.0	15.2
	Period,Comma	30.2	14.4
	Auto Period	26.1	11.8

Table 3: Entity and relation value scores for different speech recognition word error rates and various conditions of punctuation. WER=0% and All Punct are the reference conditions.

6.2. Combined Effects of WER and Automatic Punctuation

The first and last rows at each WER in Table 3, corresponding to full correct (reference) punctuation and automatic boundary detection (*Auto Period*), are plotted in Figure 2 and Figure 3 for EVS and RVS, respectively. We see from the two figures that both EVS and RVS decrease linearly as the word error rate increases. It turns out that the two linear fits in each of the two figures can be combined into a single equation by normalizing the data to the value at WER=0%. The results are the following two equations:

$$\frac{EVS}{EVS(0)} = 1 - 2WER$$

$$\frac{RVS}{RVS(0)} = 1 - 1.7WER$$

where EVS(0) and RVS(0) are the values at WER=0% for each of the *All Punct* and the *Auto Period* conditions. By including the effect of automatic period insertion, we have the following two relations:

$$\frac{EVS\_auto\_period}{EVS(ref)} = 0.865(1 - 2WER) \quad (1)$$

$$\frac{RVS\_auto\_period}{RVS(ref)} = 0.75(1 - 1.7WER) \quad (2)$$

where *EVS*(*ref*) and *RVS*(*ref*) refer to the value scores for the reference text with full punctuation.

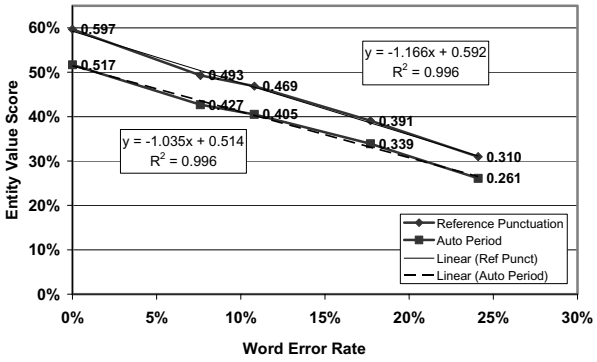

Figure 2: Entity value scores as a function of word error rate, for full reference punctuation and for automatic boundary detection (Auto Period).

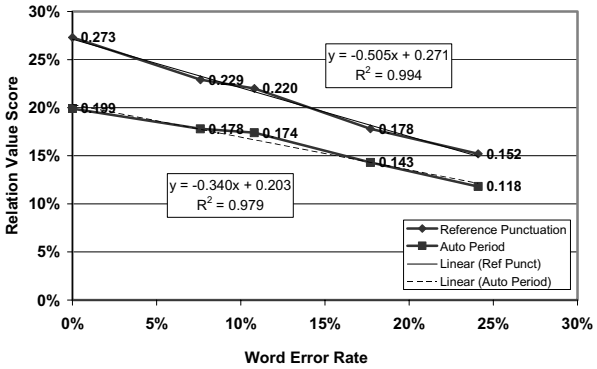

Figure 3: Relation value scores as a function of word error rate, for full reference punctuation and for automatic boundary detection (Auto Period).

These equations are of the form

$$p(1 - w * WER) \quad (3)$$

where p is a factor that includes the loss due to less than perfect punctuation (1-p is the percentage loss due to automatic punctuation), and w determines the relative loss in information extraction accuracy due to speech recognition errors.

It is clear from the above equations that, even with no speech recognition errors, automatic punctuation can result in a reduction in information extraction scores. p=0.865 corresponds to a relative reduction in EVS of 13.5% and p=0.75 corresponds to a relative reduction of RVS of 25% with automatic punctuation relative to full reference punctuation.

Equations (1) and (2) show that punctuation is relatively more important for relations than for entities (which is to be expected) but that the WER affects entities more than relations. With a WER of 25%, for example, the entity value score is reduced by 50%.

7. Conclusions

We have measured the effects of speech recognition word error rate and punctuation on information extraction. The information extraction is relatively insensitive to the correct location of sentence boundaries, but does lose performance if commas are omitted. The information extraction performance degrades linearly with increased word error rate, with a relative reduction in value scores that is approximately twice the word error rate.

8. Acknowledgements

We would like to thank LDC for recording the newswire part of the speech corpus used in this study. We also thank George Doddington for modifying the ACE scoring software to take word match into account.

9. References

[1] D. Miller, R. Schwartz, R. Stone, and R. Weischedel, "Named Entity Extraction from Broadcast News," *Proc. DARPA Broadcast News Workshop*, Herndon, VA, pp. 37-40, Feb. 28 – March 3, 1999.

[2] S. Boisen, D. Miller, R. Schwartz, R. Stone, and R. Weischedel, "Name Entity Extraction from Noisy Input: Speech and OCR," *Proc. 6th Applied Natural Language Processing Conference*, Seattle, WA, pp. 316-324, April 29-May 4, 2000.

[3] G. Doddington, A. Mitchell, M. Przybocki, L. Ramshaw, S. Strassel, and R. Weischedel, "The Automatic Content Extraction (ACE) Program - Tasks, Data, and Evaluation," *Proc. LREC,* pp. 837-840, 2004.

[4] E. Boschee, S. Bratus, S. Miller, L. Ramshaw, R. Stone, R. Weischedel, and A. Zamanian, "Experiments in Multi-Modal Automatic Content Extraction," *Proc. Human Language Technology Conference*, San Diego, CA, March 18-21, 2001.

[5] A. Stolcke, "SRILM - An Extensible Language Modeling Toolkit," *Proc. ICSLP,* Denver, Colorado, Vol. 2, pp. 901-904, September 2002.

[6] A. Stolcke and E. Shriberg, "Automatic Linguistic Segmentation of Spontaneous Speech," *Proc. ICSLP*, pp. 1005-1008, 1996.

[7] The ACE 2004 Evaluation Plan is available at http://www.nist.gov/speech/tests/ace/ace04/doc/ace04-evalplan-v7.pdf