Automatic Capitalisation Generation for Speech Input

JI-HWAN KIM & PHILIP C. WOODLAND

Cambridge University Engineering Department,
Trumpington Street, Cambridge, CB2 1PZ, UK.
{jhk23,pcw}@eng.cam.ac.uk

Abstract

Two different systems are proposed for the task of capitalisation generation. The first system is a slightly modified speech recogniser. In this system, every word in the vocabulary is duplicated: once in a decapitalised form and again in capitalised forms. In addition, the language model is re-trained on mixed case texts. The other system is based on Named Entity (NE) recognition and punctuation generation, since most capitalised words are the first words in sentences or NE words. Both systems are compared when every procedure is fully automated. The system based on NE recognition and punctuation generation shows better results by word error rate, by F-measure and by slot error rate than the system modified from the speech recogniser. This is because the latter system has a distorted language model and a sparser language model. The detailed performance of the system based on NE recognition and punctuation generation is investigated by including one or more of the following: the reference word sequences, the reference NE classes and the reference punctuation marks. The results show that this system is robust to NE recognition errors. Although most punctuation generation errors cause errors in this capitalisation generation system, the number of errors caused in capitalisation generation does not exceed the number of errors from punctuation generation. In addition, the results demonstrate that the effect of NE recognition errors is independent of the effect of punctuation generation errors for capitalisation generation.

1. Introduction

Even with no speech recognition errors, automatically transcribed speech is much harder to read due to the lack of punctuation, capitalisation and number formatting. The output format of a standard research speech recogniser is known as Standard Normalised Orthographical Representation (SNOR) (NIST, 1998a) and consists of only single-case letters without punctuation marks or numbers. The readability of speech recognition output would be greatly enhanced by generating proper capitalisation. When speech dictation is performed, the dictation system can rely on the speaker explicitly to indicate the capitalised words, although people do not want to be forced to verbally capitalise words. However, when speakers are unaware that their speech is automatically transcribed, e.g. broadcast news and conversational speech over the telephone, explicit indications of capitalised words are not given. When the input text comes from speech, the capitalisation generation task become more difficult because of corruptions of the input text caused by speech recognition errors.

The tasks of Named Entity (NE) (MUC, 1995) recognition* and enhanced speech recognition output generation are strongly related to each other, because most capitalised words are the first words in sentences or are NEs. The importance of NE recognition in automatic capitalisation was mentioned in (Gotoh, Renals & Williams, 1999). The generated punctuation and capitalisation give further clues for NE recognition. NE recognition experiments, which compare the effects of the input condition of between mixed case text and SNOR, showed that the performance deteriorates when capitalisation and punctuation information are missing (Kubala, Schwartz, Stone & Weischedel, 1998). This missing information makes certain decisions regarding proper names more difficult.

The objective of this paper is to devise automatic methods of capitalisation generation for speech input. The paper consists of seven sections. First, previous work in this area is introduced. The corpora used in the experiments are then described. Along with evaluation measures for the systems, the two different automatic capitalisation systems are presented: the first system is a slightly modified speech recogniser and the other system is based on NE recognition and punctuation generation. Finally, the detailed performance of the system based on NE recognition and punctuation generation is investigated.

2. Previous work

Many commercial implementations of automatic capitalisation are provided with word processors. In these implementations, the grammar and spelling checkers of word processors generate suggestions about capitalisation. A typical example is one of the most popular word processors, Microsoft Word. The details of its implementation was described in a U.S. patent (Rayson, Hachamovitch, Kwatinetz & Hirsch, 1998). In this implementation, whether the current word is at the start of a sentence or not was determined by a sentence capitalisation state machine. A word was defined as the text characters and any adjacent punctuation. The sentence capitalisation state machine used the characters of the current word for the transition between its possible states.

For example, if it passes a sentence ending punctuation character, the capitalisation state machine changed its state to the end punctuation state. By passing the characters of words to the capitalisation state machine, the auto correct function could determine if a particular word is at the end of a sentence, and if so, the auto correct function could determine that the next word needs to begin with an upper case letter. The capitalisation of words which are not the first word in a sentence are found by dictionary look-up. When a word is entered entirely in lower case letters, the most frequent capitalisation type of the word is assigned.

An approach to the disambiguation of capitalised words was presented in (Mikheev, 1999). The capitalised words which were located at positions where capitalisation was expected (e.g. the first word in a sentence) may be proper names or just capitalised forms of common words. The main strategy of this approach was to scan the whole of the document in order to find the unambiguous usages of words.

The importance of NE recognition in automatic capitalisation was mentioned in (Gotoh, Renals & Williams, 1999). In that study of NE tagged language models, it was stated that automatic capitalisation can possibly be achieved by programming the speech recognition decoder to produce lowercase characters apart from the capitalisation of the detected NEs. However, this is not enough for automatic capitalisation because capitalised words can normally be categorised into two groups: first words in sentences and NE words. Furthermore, some NE words are not capitalised and some non-NE words are capitalised. In addition, in some capitalised words, all characters are capitalised. Therefore, systems of automatic capitalisation have to rely on NE recognition, sentence segmentation, automatic punctuation, and the capitalisation look-up table.

NE recognition systems are generally categorised as either stochastic (typically HMM-based) or rule-based. In (Kim & Woodland, 2000b), we presented an automatic rule generating method, which uses the Brill rule inference approach (Brill, 1993, 1994), for the NE task. In (Kim & Woodland, 2000b), experimental results showed that automatic rule inference is a viable alternative to the stochastic approach to NE recognition, but it retains the advantages of a rule-based approach. In order to measure the performance of this rule-based NE recognition system, it was compared with that of IdentiFinder (Bikel, Miller & Schwartz, 1997; Kubala, Schwartz, Stone & Weischedel, 1998; Miller, Crystal, Fox, Ramshaw & Schwartz, 1997), BBN's HMM-based system which gave the best performance among the systems that participated in the 1998 Hub-4 Named Entity benchmark test (Przybocki, Fiscus, Garofolo & Pallett, 1999).

An automatic sentence segmentation method based on N-gram language modelling was described in (Stolcke & Shriberg, 1996). In their work, the performance of sentence segmentation was improved for conversational speech with the combination of other word-level features, such as POS information and turn information. The use of prosodic information combined with language cues for segmentation was pioneered in work on integrated segmentation and classification of Dialog Acts (DA, or the classification of utterances as statements, questions, agreements, etc.) in (Warnke, Kompe, Niemann & Nöth, 1998). In their approach, the optimal segmentation and classifica-

tion of DAs were searched for in the A^*-algorithm using a stochastic language model based on the word chain, a multi-layer perceptron (MLP) based on prosodic features, and a category-based language model for each DA. A combined approach for the detection of sentence boundaries and disfluencies in spontaneous speech was explained in (Stolcke, Shriberg, Bates, Ostendorf, Hakkani, Plauche, Tür & Lu, 1998). Their system combined prosodic and language model knowledge sources. The prosodic model knowledge source was modelled by decision trees, and the language model knowledge source was modelled by N-grams.

An automatic punctuation system, called Cyberpunc, which is based on only lexical information, was developed in (Beeferman, Berger & Lafferty, 1998). Their system only produced commas, under the assumption that sentence boundaries are predetermined. A method of speech recognition with punctuation generation based on both acoustic and lexical information was proposed and examined for read speech from 3 speakers in (Chen, 1999). An automatic punctuation generation method consisting of a modified speech recogniser was proposed for BN data in (Kim & Woodland, 2001). In that paper, several straightforward modifications to a conventional speech recogniser allow the system to produce punctuation and speech recognition hypotheses simultaneously. Punctuation generation for BN data was also investigated with the help of both finite state and neural-net based methods in (Christensen, Gotoh & Renals, 2001). In their work, it was shown that both methods are reasonable, and that pause duration is the strongest candidate for punctuation generation. A maximum-entropy based approach for punctuation mark annotation of spontaneous conversational speech was presented in (Huang & Zweig, 2002). Their approach viewed the insertion of punctuation as a form of tagging. Words were tagged with appropriate punctuation by a maximum entropy tagger which used both lexical and prosodic features.

3. Corpora and evaluation measures

Two different sets of data, the Broadcast News (BN) text corpus and the 100-hour Hub-4 BN data set, were available as training data for the experiments conducted in this paper. The BN text corpus (named BNText92_97 in this paper) comprises a 184 million words of BN text from the period of 1992-1997 inclusive[†]. Another set of training data, the 100-hour BN acoustic training data set released for the 1998 Hub-4 evaluation (named BNAcoustic98) consists of acoustic data and its transcription.

Broadcast News provides a good test-bed for speech recognition, because it requires systems to handle a wide range of speakers, a large vocabulary, and various domains. Three hours of test data from the NIST 1998 Hub-4 broadcast news benchmark tests were used as test data for the evaluation of the proposed system. This test data is named TestBNAcoustic98. TestBNAcoustic98 comprises 3 hours of acoustic data and the transcription. Table 1 summarises the BN training and test data. 4-gram language models were produced by interpolating language models trained on BNText92_97 and BNAcoustic98, using a perplexity minimisation method.

[Table 1]

Capitalisation types are categorised as to whether all of the characters in a word are capitalised or de-capitalised, or whether only the first character of a word is capitalised. Details of these categories are described in Table 2. Capitalised length-one words such as initials in B. B. C. are categorised as All_Cap. There are relatively few cases which are not classified as any of the categories in Table 2 (437 and 26 cases in BNAcoustic98 and TestBNAcoustic98 respectively). Most of these are surnames. For example, McWethy, MacLaine, O'Brien, LeBowe and JonBenet. All of these exceptional cases were checked manually. From this investigation, it was concluded that there is no exceptional case which cannot be treated as Fst_Cap. All of these exceptional cases were therefore classified as Fst_Cap. Table 3 shows the number of occurrences for each type of word based on the position of words in a sentence in BNAcoustic98 and TestBNAcoustic98. Table 4 shows the statistics of BNAcoustic98 and TestBNAcoustic98.

[Table 2]

[Table 3]

[Table 4]

Evaluation of a system involves scoring the automatically annotated hypothesis text against a hand annotated reference text. Scoring of text input is relatively simple because it compares capitalised words in the reference text to those in the hypothesis text, and counts the number of matched capitalised words.

However, when the input comes from speech, because of recogniser deletion, insertion and substitution errors, a straightforward comparison is no longer possible (Grishman & Sundheim, 1995). Instead, the reference and hypothesis texts must first be automatically aligned. This is a complex process and involves attempting to determine which part of recogniser output corresponds to which part of the transcript.

Once the alignment is completed, correct/incorrect decisions for all the capitalised words can be made. We define the symbols as C for the number of correct capitalised words, S for the number of substitution errors, D for the number of deletion errors, I for the number of insertion errors, N for the number of capitalised words in reference, and M for the number of capitalised words in hypothesis. From the above definitions, it is clear that $N = C + S + D$ and $M = C + S + I$.

Two important metrics for assessing the performance of an information extraction system are *recall* and *precision*. These terms are borrowed from the information retrieval community. Recall (R) refers to how much of the information that should have been extracted was actually correctly extracted. Precision (P) refers to the reliability of the information extracted. These quantities are defined as:

$$P = \frac{\text{number of correct capitalised words}}{\text{number of capitalised words in hypothesis}} = \frac{C}{M} \tag{1}$$

and

$$R = \frac{\text{number of correct capitalised words}}{\text{number of capitalised words in reference}} = \frac{C}{N} \tag{2}$$

Although theoretically independent, in practice recall and precision tend to operate in trade-off relationships. An attempt to increase a recall frequently compromises precision. Likewise, the optimisation of precision is often detrimental to recall.

The F-measure (Makhoul, Kubala, Schwartz & Weischedel, 1999) is the uniformly weighted harmonic mean of precision and recall:

$$F = \frac{RP}{(R+P)/2} = \frac{2C}{N+M} \tag{3}$$

Another evaluation metric called Slot Error Rate (SER) was defined in (Makhoul, Kubala, Schwartz & Weischedel, 1999) as follows:

$$\text{SER} = \frac{\text{number of capitalisation generation errors}}{\text{number of capitalised words in reference}} = \frac{S+D+I}{C+S+D} \tag{4}$$

The difference between SER and $(1-F)$ is the weight given to D and I. The value of $(1-F)$ is calculated as:

$$(1-F) = \frac{N+M-2C}{N+M} = \frac{S+(D+I)/2}{(N+M)/2} = \frac{S+(D+I)/2}{C+S+(D+I)/2} \tag{5}$$

To implement scoring, version 0.7 of the NIST Hub-4 IE scoring pipeline package (NIST, 1998b) was used. Although this scoring pipeline was developed for the NE recognition system evaluation only, this scoring pipeline can be applied for the evaluation of a capitalisation generation system by small manipulations of the reference and the hypothesis files.[‡]This pipeline package aligns the reference and the hypothesis files first. It then calculates scores based on how well the capitalisation types of the capitalised words (All_Cap and Fst_Cap) in the reference file agree with those in the hypothesis file. In the scoring definition used for evaluation on NE recognition systems, a half score is given for words whose capitalisation types are All_Cap in the reference file and Fst_Cap in the hypothesis files, and Fst_Cap in the reference file and All_Cap in the hypothesis files.

4. Automatic capitalisation generation

In this section, two different automatic capitalisation generation systems are presented. The first system is a slightly modified speech recogniser. In this system, every word in its vocabulary is duplicated: once in a de-capitalised form and again in the two capitalised forms. In addition, its language model is re-trained on mixed case texts. The other system is based on NE recognition and punctuation generation, since most capitalised words are first words in sentences or NE words.

4.1. Automatic capitalisation generation by modifications of speech recogniser

The method of automatic capitalisation generation presented in this section is a slightly modified form of a conventional speech recogniser. As the aim of speech recognition is to find only the best word sequence for the given speech signal, speech recognition systems do not normally recognise the capitalisation of words. Therefore, the words registered in a vocabulary and a pronunciation dictionary are not case-sensitive in a conventional speech recognition system. In addition, it is not necessary to train language models of this system on case sensitive texts.

Small modifications to a conventional speech recognition system, however, can produce case sensitive outputs. The following three modifications are required:

1. Every word in its vocabulary is duplicated for the three different capitalisation types (All_Cap, Fst_Cap, No_Cap).
2. Every word in the pronunciation dictionary is duplicated as for the vocabulary duplication with all duplicates having the same pronunciation.
3. The Language Model (LM) is re-trained on mixed case texts.

[Figure 1]

Figure 1 illustrates the overall capitalisation generation system which is modified from a conventional speech recognition system. As the LM is trained on case sensitive training data, this LM is sparser than that used by the conventional speech recogniser. The same acoustic score will be assigned to duplicated words, since they have the same pronunciations. However, different hypotheses will be generated using the different LM scores. Speech recognition is performed, and the best hypothesis which includes capitalisation is generated.

As sentence boundary information is necessary to generate capitalisation for the first word of a sentence, the capitalisation generation system also has two modifications to a conventional speech recognition system to allow it to generate punctuation marks. First, the pronunciation of punctuation marks is registered as silence in the pronunciation dictionary. Secondly, the LM is trained on mixed-case texts which contain punctuation marks.

The correlation between punctuation and pauses was investigated in (Chen, 1999).

That experiment showed that pauses closely correspond to punctuations. The correlation between pause lengths and sentence boundary marks was studied for BN data in (Gotoh & Renals, 2000). In that study, it was observed that the longer the pause duration, the greater the chance of a sentence boundary existing. The pause duration and other prosodic features were examined on the punctuation generation for BN data in (Christensen, Gotoh & Renals, 2001). In their work, it was shown that pause duration is the strongest candidate for punctuation generation. Pause duration information was also used in the punctuation annotation of spontaneous conversational speech using a maximum entropy tagger with the help of lexical information in (Huang & Zweig, 2002). For the detection of sentence boundaries and disfluencies in spontaneous speech, studied in (Stolcke, Shriberg, Bates, Ostendorf, Hakkani, Plauche, Tür & Lu, 1998), an N-gram, which included turn and pause information, outperformed N-gram which did not have this information.

Although some instances of punctuation do not occur at pauses, it is convenient to assume that the acoustic pronunciation of punctuation is silence. The details of our punctuation generation system were described in (Kim & Woodland, 2001).

4.2. Automatic capitalisation generation based on NE recognition and punctuation generation

The method of capitalisation generation presented in this section is based on NE recognition and punctuation generation, since most capitalised words are either the first words in sentences or NE words. This method uses the rule-based (transformation-based) NE recognition system (Kim & Woodland, 2000b), which uses the Brill rule inference approach (Brill, 1993), and the punctuation generation system which incorporates prosodic information along with acoustic and language model information (Kim & Woodland, 2001).

4.2.1. Description of the NE recognition system used

For NE recognition, the learning procedure begins by using an unannotated input text. For all words whose NE classes and NE boundaries are incorrect, the rules to recognise these NE classes and NE boundaries correctly are generated according to their appropriate rule templates. At each stage of learning, the learner finds the transformation rules which when applied to the corpus result in the best improvement. The improvement can be calculated by comparing the current NE tags after the rule is applied with the reference tags. After finding this rule, it is stored and applied in order to change the current tags. This procedure continues until no more transformations can be found.

For example, the learning procedure begins with using the following unannotated input text in SNOR form.

MR MANDELSON HAD MADE CLEAR FOR THE FIRST TIME THAT ALL THE NEW

INSTITUTION INCLUDING THE VARIOUS CROSSBORDER BODIES CREATED YESTERDAY ...

The NE classes of 'MANDELSON' and 'YESTERDAY' in the example text are incorrect. If the rule, "if the current word is 'MR', then change the NE class of the next word to 'PERSON"', results in the best improvement over the whole of the corpus, this rule is applied, and the example text changes as follows:

MR <ENAMEX TYPE="PERSON">MANDELSON</ENAMEX> HAD MADE CLEAR FOR THE FIRST TIME THAT ALL THE NEW INSTITUTION INCLUDING THE VARIOUS CROSSBORDER BODIES CREATED YESTERDAY ...

The same procedure continues using the rule, "if the current word is 'YESTERDAY', then change the NE class of the current word to 'DATE"'. After this rule is applied, the example text changes as follows:

MR <ENAMEX TYPE="PERSON">MANDELSON</ENAMEX> HAD MADE CLEAR FOR THE FIRST TIME THAT ALL THE NEW INSTITUTION INCLUDING THE VARIOUS CROSSBORDER BODIES CREATED <TIMEX TYPE="DATE">YESTERDAY</TIMEX> ...

The two rules which give the largest improvements when the training procedure starts in (Kim & Woodland, 2000a) are as follows:

1. If the current word is 'DOLLARS' and the feature of the previous word is 'NUMERIC', then change the word classes of the current and previous words to 'MONEY'.
2. If the current word is 'NINETEEN' and the feature of the current word is 'NUMERIC', then change the word class of the current word to 'DATE'

In testing, the rules are applied to the input text one-by-one according to a given order. If the conditions for a rule are met, then the rule is triggered and the NE classes of the words are changed if necessary.

In (Kim & Woodland, 2000b), the performance of the rule-based NE recognition system was compared with BBN's commercial stochastic NE recogniser called IdentiFinder. For the baseline case (SNOR), both systems show almost equal performance, and are also similar when additional information such as punctuation, capitalisation and name lists is given. When input texts are corrupted by speech recognition errors, the performance of both systems are degraded by almost the same amount. Although the rule-based approach is different from the stochastic method, which is recognised as one of the most successful methods, the rule-based system shows the same level of performance.

4.2.2. Description of the punctuation generation system used

Punctuation generation uses two straightforward modifications of a conventional speech recogniser described in Section 4.1. First, the pronunciation of punctuation marks is registered as silence in the pronunciation dictionary. Secondly, the language model is trained on the texts which contain punctuation marks. These modifications allow the system to produce punctuation and speech recognition hypotheses simultaneously.

Multiple hypotheses are produced by the automatic speech recogniser and are then re-scored using a prosodic feature model based on Classification And Regression Trees (CART) (Breiman, Friedman, Olshen & Stone, 1983). A set of 10 prosodic features were used for punctuation generation. When prosodic information is incorporated, the F-measure was improved by 19% relative. At the same time, small reductions in word error rate were obtained.

4.2.3. Procedures for capitalisation generation

[Figure 2]

Figure 2 shows the procedure applied by the capitalisation generation system based on NE recognition and punctuation generation. As shown in Figure 2, the capitalisation generation system proposed in this section consists of 8 steps. The various stages shown in Figure 2 are explained below.

The simplest method of capitalisation generation is to capitalise the first characters of words which are the first words in sentences and the first characters of NE words whose NE classes are 'ORGANIZATION', 'PERSON', or 'LOCATION', followed by capitalisation of initials. These straightforward processes are performed from steps 1 to 4 in Figure 2.

The results of capitalisation generation can be improved by using frequency of occurrence of NE words in the training texts. Some NE words are used in de-capitalised forms and some non-NE words are used in capitalised forms. Also, all characters should be capitalised in some first words in sentences. Many of these capitalisation types are corrected by look-up in a frequency table of words based on NE classes. This information is used in steps 5, 6, and 7. In step 5, the most frequent capitalisation type within an NE class is given to NE words which are not the first word in a sentence. In step 6, the same process is applied to non-NE words which are not the first word in a sentence. In step 7, if a word with the 'ORGANIZATION' class is the first word in a sentence, and its most frequent capitalisation type is All_Cap, then the capitalisation type of this word is changed to All_Cap.

Further improvement can be achieved by using context information to dis-ambiguate the capitalisation types of words which have more than one capitalisation type such as the word 'bill' (which can be used as a person's name as well as a statement of account). The context information about capitalisation generation is encoded in a set of simple rules rather than the large tables of statistics used in stochastic methods.

The transformation-based approach used in the development of the rule-based NE recognition system described in Section 4.2.1 is applied in the automatic generation of these rules for capitalisation generation. The automatic capitalisation generation views finding the capitalisation types of words as a form of tagging the NEs. Applicable rules are generated according to the rule templates in the transformation-based approach. Because each capitalised word is treated as one entity in the capitalisation generation, boundary expansion rule templates are not considered in the design of rule templates.

[Table 5]

Six rule templates are used for the generation of bigram rules for capitalisation generation. These six rule templates are shown in Table 5. The rule templates consist of pairs of characters and a subscript, and w, t, c denote that templates are related to words, NE classes and

capitalisation types respectively. Subscripts show the relative distance from the current word, e.g. 0 refers the current word.

For these rules, the range of rule application is set to be the current word only, because each capitalised word is treated as one entity. For example, if an applicable rule $R(n_0, n_1; n_2)$ is generated by the rule template (w_0, w_1), $R(n_0, n_1; n_2)$ means "change the capitalisation type of the current word to the capitalisation type n_2, when the current word is n_0 and the following word is n_1". Similarly, $R(n_3, n_4; n_5)$ generated by the rule template (w_0, c_1) means "change the capitalisation type of the current word to the capitalisation type n_5, when the current word is n_3 and the capitalisation type of the following word is n_4".

A particular problem is the effect of words encountered in the test data which have not been seen in the training data. One way of tackling the situation is to build separate rules for unknown words. The training data are divided into two groups. If words in one group are not seen in the other group, these words are regarded as unknown words. The same rule generation procedures are then applied. The bigram rules generated from 6 rule templates described in Table 5 are applied one-by-one in step 8 according to a given order.

5. Experimental results

There are two different systems of generating capitalisation: a system modified from a speech recogniser (named ModSR) and a system based on NE recognition and punctuation generation (named NEPuncBased).

NEPuncBased uses the rule-based NE recognition system in (Kim & Woodland, 2000b), which generates rules automatically, and the punctuation generation system in (Kim & Woodland, 2001), which incorporates prosodic information along with acoustic and language model information.

A version of the HTK system (Woodland, 2002) for Broadcast News (BN) transcription running under 10 times real time (10xRT) (Odell, Woodland & Hain, 1999) was used for both capitalisation generation systems. The first step of the system is a segmentation stage which converts the continuous input stream into segments with the aim of each segment containing data from a single speaker and a single audio type. Each segment is labelled as being either a wide-band or narrow-bandwidth signal.

The actual recogniser runs in two passes which both use cross-word triphone decision-tree state clustered HMMs with Gaussian mixture output distributions and a N-gram language model. The first pass uses gender-independent (but bandwidth-specific) HMMs with a 60k trigram language model to get an initial transcription for each segment. This transcription is used to determine the gender label for the speaker in each segment by alignment with gender-dependent HMMs. Sets of segments with the same gender/bandwidth labels are clustered for unsupervised Maximum Likelihood Linear Regression (MLLR) (Leggetter & Woodland, 1995) adaptation. The MLLR transforms for each set of clustered segments are computed using the initial transcriptions of the segments and the gender-dependent HMMs used for the second pass. The adapted HMMs along with a 4-gram language model is used in the second stage of decoding and produces the final output.

Implementation details of the HTK BN transcription system (with few constraints on computing power) were given in (Woodland, Hain, Johnson, Niesler, Whittaker & Young, 1998; Woodland, Hain, Moore, Niesler, Povey, Tuerk & Whittaker, 1999), and those of the HTK

10xRT BN transcription system were described in (Odell, Woodland & Hain, 1999). In order to speed up the full system, the 10xRT system uses simpler acoustic models and a simplified decoding strategy.

Using the HTK 10xRT system, speech recognition is performed first for TestBNAcoustic98. As punctuation and capitalisation are not considered at this stage, the test condition is the same as for the NIST 1998 Hub-4 broadcast news benchmark tests. The Word Error Rate (WER) of the speech recogniser was measured as 16.7%.

The HTK 10xRT BN transcription system reported 16.1% of overall WER for the NIST 1998 Hub-4 BN benchmark test (Pallett, Fiscus, Garofolo, Martin & Przybocki, 1999). The difference between the reported performance in (Pallett, Fiscus, Garofolo, Martin & Przybocki, 1999) and the performance measured in this paper is 0.6%. The system used in this paper differs from the HTK 10xRT system used in the 1998 Hub-4 BN benchmark test in four aspects: the absence of a category-based language model (Niesler, Whittaker & Woodland, 1998), the amount of language model training data, the difference in vocabulary size, and the absence of the procedure to obtain more precise word start and end time information.

The results of both systems are compared for a speech recognition output on the basis that every capitalisation procedure is fully automated. Then, the performance of NEPuncBased is investigated with additional information: reference word sequences, reference NE classes and reference punctuation marks. As NEPuncBased follows the 8 steps described in Figure 2, the effect of each step is examined when reference word sequences, reference NE classes, and reference punctuation marks are provided. The effect of each step is also examined for speech recognition output when every capitalisation procedure is fully automated.

5.1. Results: The system modified from a speech recogniser (ModSR)

The first automatic capitalisation generation system is implemented by small modifications to the HTK Broadcast News (BN) transcription system. First, every word in the pronunciation dictionary of the HTK system is duplicated with its pronunciation into the three different capitalisation types (All_Cap, Fst_Cap, and No_Cap). Second, the language model is re-trained on mixed case transcriptions of BNText92_97 and BNAcoustic98.

Table 6 shows the results of capitalisation generation for TestBNAcoustic98 using this system. When WER is measured, words are changed into single case forms from the reference and hypothesis in order to measure the pure speech recognition rate. As the speech recognition output contains punctuation marks, WER″, which is the WER after punctuation marks are removed and words are changed to single case, is introduced.

[Table 6]

For punctuation generation, the HTK system gave 22.73% WER in (Kim & Woodland, 2001). The difference between WER in punctuation generation and that in capitalisation generation is 0.24%. The degradation is caused by the introduction of an increased size of vocabulary and pronunciation dictionary. The performance degradations can be analysed as follows:

1. LM distorted by first words of sentences

   In many cases, the first word of a sentence is not an NE. Most of these words are not

capitalised, if they are used in the middle of sentences. As there are 1,873 sentences in TestBNAcoustic98, the average number of words in a sentence in TestBNAcoustic98 is 16.9 words. Among the first words in sentences, 91.3% of these words are not NEs. Therefore, approximately, 5.4% ((1/16.9) $\times$ 0.913) of counted word sequences are wrong, because a capitalised word and a de-capitalised word should be regarded as different words even if they have the same character sequence.

2. Sparser LM

   Due to the limited amount of training data, many of the possible word sequences in the test data are not observed in the training data. As the size of the vocabulary is increased, LMs are sparser and estimating probabilities of word sequences becomes more difficult. The HTK system generates initial hypotheses using trigram language models and re-scores these hypotheses using 4-gram language models. As the size of vocabulary is multiplied by three, these LMs are sparser and the search space is increased.

In addition to the effects for capitalisation generation, caused by the two factors of speech recognition degradation, loss of half scores in the evaluation of capitalisation generation affects the performance. If NE recognition and capitalisation generation are performed as post-processing of speech recognition, it is possible to obtain half scores for the words which are mis-recognised in speech recognition but are located next to NE signalling words.

5.2. Results: The system based on NE recognition and punctuation generation (NEPuncBased)

The steps of the capitalisation generation system depicted in Figure 2 start from the single case speech recognition output with punctuation marks and NE classes. In this system, multiple hypotheses which include punctuation marks are produced by the HTK system and are re-scored by prosodic information. Then NE recognition is performed for this speech recognition output. Capitalisation generation follows this speech recognition output with generated NE classes.

The results of automatic punctuation generation according to various scale factors to the prosodic feature model were presented in (Kim & Woodland, 2001). The scale factor to prosodic feature model is set to be 0.71 at which WER is minimised. This automatic punctuation generation system gave WER, F-measure and SER as 22.55%, 0.4239 and 83.36% respectively for TestBNAcoustic98. Further details of this prosody combined system for punctuation generation and speech recognition were given in (Kim & Woodland, 2001).

NE recognition is performed for the best re-scored hypothesis. As an NE recogniser, the rule-based NE recogniser trained under the condition of 'with punctuation and name lists but without capitalisation' is used. This NE recogniser reported an F-measure of 0.9007 and an SER of 16.68% in (Kim & Woodland, 2000b) for the reference transcription of TestBNAcoustic98. More details of this NE recogniser were discussed in (Kim & Woodland, 2000b).

The frequency table and bigram rules were constructed using the transcription of BNAcoustic98. Table 7 shows the result of capitalisation generation based on NE recognition and punctuation generation. As this system does not increase the size of the vocabulary, there is no degradation in WER. Compared to the other capitalisation generation system (ModSR), this system (NEPuncBased) shows better results by: 0.42% in WER, 0.41% in WER$''$, 2.62% in SER,

and 0.0089 in F-measure. The factors which cause these differences were explained as 'the distortion of LM', 'sparser LM', and 'loss of half scores' in Section 5.1.

[Table 7]

6. Analysis of performance of the system based on NE recognition and punctuation generation (NEPuncBased)

The effects of speech recognition errors, NE recognition errors and punctuation generation errors are accumulated in the results of NEPuncBased in Table 7. In this section, the performance of NEPuncBased is investigated by including one or more of the following: reference word sequences, reference NE classes and reference punctuation marks. The total effect of the accumulated errors is examined, and the contribution of each step in NEPuncBased is tested for reference word sequences, NE classes and punctuation marks. The effect of each step is also examined for speech recognition output when every capitalisation procedure is fully automated. Then, the effects of speech recognition and punctuation generation errors are examined. The performance of NEPuncBased is compared with that of Microsoft Word 2000 for a reference text in order to remove the effect of speech recognition errors.

6.1. The contribution of each experimental step

In order to measure the pure contribution from each step in the capitalisation generation system based on NE classes and punctuation marks, the contributions were measured first for reference word sequence, reference NE classes and reference punctuation marks.

Table 8 shows the result of the capitalisation generation system based on NE classes and punctuation marks for these test conditions. The F-measure is measured as 0.9756 and the SER as 4.89%. After removing the effects of speech recognition errors, NE recognition errors and punctuation generation errors, the F-measure is improved by 0.2350 (0.9756 - 0.7406) and the SER by 41.04% (45.93 - 4.89).

[Table 8]

Table 9 shows the capitalisation generation results with different combinations of experimental steps. By just performing step 1 (the first character of the first word in each sentence is capitalised), an F-measure of 0.5494 is already obtained, although the recall (0.3814) is quite poor compared to the precision (0.9818). By performing step 2, in addition to step 1, the F-measure is increased to 0.8448.

Adding steps 3 and 4, which can be done by straightforward processes without the need for training data, an F-measure of 0.9247 is obtained for capitalisation generation. With steps 5, 6 and 7 which depend on the use of frequency tables, the result can be increased to 0.9694. This is increased to 0.9756 points in F-measure using bigram rules. Table 9 shows these results.

[Table 9]

The contribution of each step was also measured for speech recognition output when every capitalisation procedure is fully automated. Table 10 shows these results. By just performing step 2, in addition to step 1, an F-measure of 0.6194 is obtained, although the recall (0.5308) is poor compared to the precision (0.7434). By adding steps 3 and 4, an F-measure of 0.6799 is obtained for capitalisation generation. With steps 5, 6 and 7, the F-measure increases to 0.7339. Using bigram rules, this increases to 0.7406 points in F-measure.

[Table 10]

6.1.1. Analysis: The result of capitalisation generation when reference word sequences, NE classes and punctuation marks are provided

The results of capitalisation generation are analysed for reference word sequences when NE classes and punctuation marks are also provided because these results do not have any type of recognition error apart from capitalisation generation errors. The capitalisation generation system based on NE classes and punctuation marks reports an F-measure of 0.9756 with 236 errors for TestBNAcoustic98 when reference word sequences, punctuation marks and NE classes are provided. These 236 errors can be categorised into the following three groups:

1. Errors due to the inconsistency of capitalisation (Group 1)
2. Errors due to limited number of observations in training data (Group 2)
3. Errors not included in Group 1 and Group 2 (Group 3)

Groups 1 and 2 are not totally exclusive of each other. The number of errors in Group 1 can be measured by substituting the training data with the test data and repeating the experiment. After this substitution, there were still 100 errors with an F-measure of 0.9896. These 100 errors were examined manually. Most of them are caused by inconsistency of capitalisation which cannot be corrected by bigrams. For example:

- News in “Lisa Stark, A. B. C. News, Washington” (normally A. B. C. news)
- the President (normally the president apart from the President of U. S. A.)
- World Today (programme name)
- South, East .... (normally south, east but sometimes capitalised in weather forecast)
- Main Street in “U. S. props up Japan’s currency from Wall Street to Main Street” (normally main street)

The errors in Group 2 show that they can be corrected if the size of the training data is increased. Assume that a word in test data is observed enough times for correct modelling if it is observed in training data more than twice (≥ 3) with its NE class and its capitalisation type. On this assumption, capitalisation errors in Group 2 can be categorised into the following 4 sub-categories:

1. Errors at an unknown word (Group 2-1)
2. Errors at a word never seen in the training data with its NE class (Group 2-2)
3. Errors at a word seen only once in the training data with its NE class (Group 2-3)

4. Errors at a word seen twice in the training data with its NE class (Group 2-4)

Among 236 total errors, the number of errors in Group 2-1, 2-2, 2-3 and 2-4 are counted as 25, 23, 9 and 0 respectively. These numbers constitute 24.15% of total errors.

A word which has a capitalisation type error in Group 3 is observed frequently enough with its NE class. As these errors are not caused by the inconsistency in capitalisation, corrections for these errors are difficult using the current method of capitalisation generation.

Among these three categories of errors in capitalisation generation, only the errors in Group 2 can be corrected if the size of the training data is increased. The errors in Group 2 consist of 25.85% of total errors and the F-measure of the system on the current input condition is 0.9756. If the errors in Group 2 are corrected, the F-measure of this capitalisation generation system is expected to be increased to 0.9819 (0.9756 + (1-0.9756) $\times$ 0.2585).

It is currently believed that an F-measure of 0.9756 in capitalisation generation on the condition of reference word sequences, punctuation marks and NE classes is a good result given the relatively small amount of training data i.e. only BNAcoustic98 was used for the construction of the frequency table and the bigram rules.

6.2. The effect of NE recognition errors

In order to measure the effect of NE recognition errors in the capitalisation generation system based on NE classes and punctuation marks, the results of capitalisation generation are examined for reference word sequences and reference punctuation marks. However, NE classes are generated by an NE recogniser.

Table 11 shows the results of capitalisation generation for reference word sequences, generated NE classes and reference punctuation marks. As the F-measure of capitalisation generation for reference word sequences, NE classes and punctuation marks was measured as 0.9756, the effect of NE recognition errors on capitalisation generation is measured with a degradation in F-measure of 0.0158 (0.9756 - 0.9585). The degradation in SER is measured as 3.20%.

[Table 11]

6.2.1. Analysis: The effect of NE recognition errors

Steps 2, 5, 6 and 7 of the capitalisation generation system described in Figure 2 are based on NE classes. In this section, the effect of NE recognition errors for the overall performance of capitalisation generation is analysed.

The statistics of TestBNAcoustic98 were shown in Tables 3 and 4. According to these tables, the number of initial words which are NEs is 543 and the number of NE words which are first words in sentences and which have a capitalised first character is 143. Among NE words, these 543 initials and 143 NEs at the beginning of sentences can be capitalised correctly without the help of the NE recognition system. As the total number of NEs in TestBNAcoustic98 is 3,149, the number of NEs which require the help of the NE recognition system is roughly 2,463 (3,149 - 543 - 143).

As the F-measure of the NE recogniser is 0.9007 for NE recognition, the capitalisation of

about 245 (2,463 $\times$ (1 - 0.9007)) NE words may be affected by the NE recognition errors. This number of words constitutes 5.1% of the total capitalised words. However, the actual degradation caused by the errors in NE recognition is measured as 0.0158 of F-measure. This implies that this capitalisation generation system is robust to NE recognition errors.

6.3. The effect of punctuation generation errors

In order to measure the effect of punctuation generation errors in the capitalisation generation system based on NE classes and punctuation marks, the results of capitalisation generation are examined for the reference word sequences, reference NE classes and generated punctuation marks. The punctuation generation system using combined information of an LM and a prosodic feature model is used. It generates punctuation marks with an F-measure of 0.7830 and an SER of 32.30% for the reference transcription of TestBNAcoustic98. More details of this punctuation generation system were given in (Kim & Woodland, 2001).

Table 12 shows the result of capitalisation generation for reference word sequences, reference NE classes and generated punctuation marks. As the F-measure of capitalisation generation for reference word sequences, NE classes and punctuation marks was measured as 0.9756, the effect of punctuation generation errors on capitalisation generation is measured as an F-measure of 0.0909 (0.9756 - 0.8847). The degradation in SER is measured as 18.21%.

[Table 12]

6.3.1. Analysis: The effect of punctuation generation errors

Steps 1, 5, 6 and 7 of the capitalisation generation system depicted in Figure 2 are based on punctuation marks. According to the statistics of TestBNAcoustic98 shown in Tables 3 and 4, the number of non-NE words which have a capitalised first character and which are first words in sentences is 1,603.

Punctuation marks whose place is correct but type is wrong are meaningful in punctuation generation and obtain half scores. However, punctuation type errors between commas and full stops, and between commas and question marks are not meaningful for capitalisation generation, because the words next to commas are normally de-capitalised. If the half scores are given in punctuation generation only between full stops and question marks, the F-measure of punctuation generation decreases to 0.6826.

The maximum number of words whose capitalisation types are possibly affected by punctuation generation errors can be roughly estimated as 1,603 $\times$ (1 - 0.6826) = 509. This number of words constitute 10.56% of the total number of capitalised words. The actual degradation caused by punctuation generation errors is measured as an F-measure of 0.0909. This implies that most punctuation generation errors cause errors in capitalisation generation, but the number of errors caused in capitalisation generation do not exceed the number of errors in punctuation generation.

6.4. The correlation between the effects of NE recognition errors and the effects of punctuation generation errors

In this section, the correlation between the effects of NE recognition errors and those of punctuation generation errors to capitalisation generation are examined. NE recognition and punctuation generation are performed for the reference transcription of TestBNAcoustic98, in which every word is de-capitalised and every punctuation mark is removed. The rule-based NE recogniser and the punctuation generation system, which uses the combined information of an LM and a prosodic feature model, are used.

Using these NE recogniser and punctuation generation systems, punctuation marks are produced first for the transcription of TestBNAcoustic98, then NE recognition is performed for the reference transcription with these generated punctuation marks. The capitalisation generation is carried out for this result of NE recognition and punctuation generation for the transcription of TestBNAcoustic98.

Table 13 shows the results of capitalisation generation for NE recognition and punctuation generation output from reference word sequences. The simultaneous effects of NE recognition errors and punctuation generation errors on capitalisation generation are measured as a degradation in F-measure of 0.1065 and in SER of 21.36%. As the effect of NE recognition errors on capitalisation generation and the effect of punctuation generation errors on capitalisation generation are measured as 0.0158 and 0.0909 in F-measure respectively (3.20% and 18.21% in SER respectively), it is shown that these simultaneous effects are almost equivalent to the sum of individual effects. This suggests that the effect of NE recognition errors is independent of the effect of punctuation generation errors for capitalisation generation.

[Table 13]

6.5. Comparison with Microsoft Word 2000

An experiment of automatic capitalisation generation was conducted using Microsoft Word 2000 for the first 10.7% words of TestBNAcoustic98 (3882 words, 468 of which are capitalised). As it provides suggestions about both grammar and spelling, its suggestions are checked manually and only suggestions regarding capitalisations are accepted.

According to the description in (Rayson, Hachamovitch, Kwatinetz & Hirsch, 1998) and its capitalisation generation output for the part of TestBNAcoustic98, capitalisation of words which are not first words in sentences seems to be performed by dictionary look-up. When a word is entered in all lower case, the capitalisation is applied for the word to have the greatest consistency in matching the capitalisation.

With this dictionary look-up method, ambiguous words such as ‘bill’ cannot be dis-ambiguated. In a sentence like “President bill Clinton says”, ‘bill’ should be capitalised: the error occurs because the word ‘bill’ is more frequently used as a statement of account in a de-capitalised form rather than a person’s name. Dis-ambiguation of the capitalisation type of words which can have more than one type can be achieved by using context information.

The performance of NEPuncBased was compared to that of Microsoft 2000 for the same part of TestBNAcoustic98. As the reference sequence of words and punctuation marks were given as input when automatic capitalisation generation was performed by Microsoft Word

2000, capitalisation is generated by NEPuncBased for the reference word sequences, generated NE classes and reference punctuation marks. Table 14 shows the results of capitalisation generation by NEPuncBased for the first 10.7% words of TestBNAcoustic98. Compared to Microsoft, NEPuncBased shows better results by 0.0687 in F-measure and by 11.62% in SER.

[Table 14]

6.6. Estimation: Results of the system based on NE recognition and punctuation generation when every procedure is fully automated

In Section 5.2, the capitalisation generation system based on NE recognition and punctuation generation reported an F-measure of 0.7406. In this section, this result is compared with the results expected from the previous conclusions: the performance of NE recognition is degraded linearly according to speech recognition errors (Kim & Woodland, 2000b), and the effect of NE recognition errors is independent of the effect of punctuation generation errors for capitalisation generation (Section 6.4).

The experiment in Section 5.2 used a punctuation generation system which reported an F-measure of 0.4239 at a scale factor of 0.71 and reported 16.86% of WER′ (WER after removing punctuation marks from a reference and a hypothesis) at this scale factor. In addition to this punctuation generation system, the experiment used an NE recognition system which reported an F-measure of 0.9007. Since an experiment in (Kim & Woodland, 2000b) reported that the performance of an NE recogniser is linearly degraded by 0.0062 points in F-measure per 1% of additional WER, the capitalisation generation system based on NE recognition and punctuation generation is expected to obtain the F-measure of 0.7962 (0.9007 - 0.0062 $\times$ 16.86) for NE recognition.

As shown in Section 6.2, the result of capitalisation generation is degraded by an F-measure of 0.0158 due to NE recognition error of an F-measure of 0.0993 (1 - 0.9007). The degradation of capitalisation generation caused by NE recognition errors (assuming that this degradation is proportional to NE recognition errors) is expected to be 0.0324 (0.0158 $\times$ (1-0.7962)/0.0993).

As shown in Section 6.3, the result of capitalisation generation is degraded by an F-measure of 0.0909 due to punctuation generation errors of an F-measure of 0.2170 (1 - 0.7830). The degradation of capitalisation generation caused by punctuation generation errors (assuming that this degradation is proportional to punctuation generation errors) is expected to be 0.2391 (0.0909 $\times$ (1-0.4292)/0.2170).

If it is assumed that the effect of NE recognition errors is independent of the effect of punctuation generation errors for capitalisation generation, the total degradation of capitalisation generation caused by NE recognition errors and punctuation generation errors is expected to be 0.7041 (0.9756 - 0.0324 - 0.2391).

Based on this expectation, the result of capitalisation generation of an F-measure of 0.7406 is believed to be a reasonable result when every procedure is fully automated.

7. Conclusions

In this paper, an important area of transcription readability improvement, automatic capitalisation generation, has been discussed. Two different systems have been proposed for this task. The first is a slightly modified speech recogniser. In this system, every word in its vocabulary is duplicated: once in a de-capitalised form and again in capitalised forms. In addition, the language model is re-trained on mixed case texts. The other system is based on NE recognition and punctuation generation since most capitalised words are first words in sentences or NE words.

In order to compare the performance of the proposed systems, experiments of automatic capitalisation generation were performed for 3 hours of broadcast news test data (TestBNAcoustic98). The results of both systems have been compared on the basis that every procedure is fully automated. The system based on NE recognition and punctuation generation showed better results in WER, in F-measure and in SER than the system modified from the speech recogniser, because the latter system has a distorted LM and a sparser LM.

The system based on NE recognition and punctuation generation follows the 8 steps described in Figure 2. The effect of each step was examined first when reference word sequences, reference NE classes, and reference punctuation marks were provided. More than 0.92 points of F-measure for capitalisation were obtained using straightforward steps without training data. The effect of each step was also measured for speech recognition output when every capitalisation procedure was fully automated. About 0.68 points of F-measure were measured using straightforward steps without training data.

The performance of the system based on NE recognition and punctuation generation has been investigated for the additional clues: reference word sequences, reference NE classes and reference punctuation marks. The results showed that this system is robust to NE recognition errors. Although most punctuation generation errors cause errors in this capitalisation generation system, the number of errors caused in capitalisation generation does not exceed the number of errors in punctuation generation. In addition, the results demonstrate that the effect of NE recognition errors is independent of the effect of punctuation generation errors for capitalisation generation.

8. Acknowledgements

Ji-Hwan Kim acknowledges the support of the British Council, LG company, GCHQ and the EU Coretex project.

Footnote

*The Named Entity (NE) task requires the recognition of named entities (names of locations, persons and organisations), temporal expressions (dates and times) and numerical expressions (monetary amounts and percentages).
†The 1992-1996 part was provided by the LDC and the 1997 part by Primary Source Media.
‡Surround the words whose capitalisation types are All_Cap by the “ORGANIZATION” NE class starting and end tags and enclose the words whose types are Fst_Cap by the “PERSON” NE class tags.

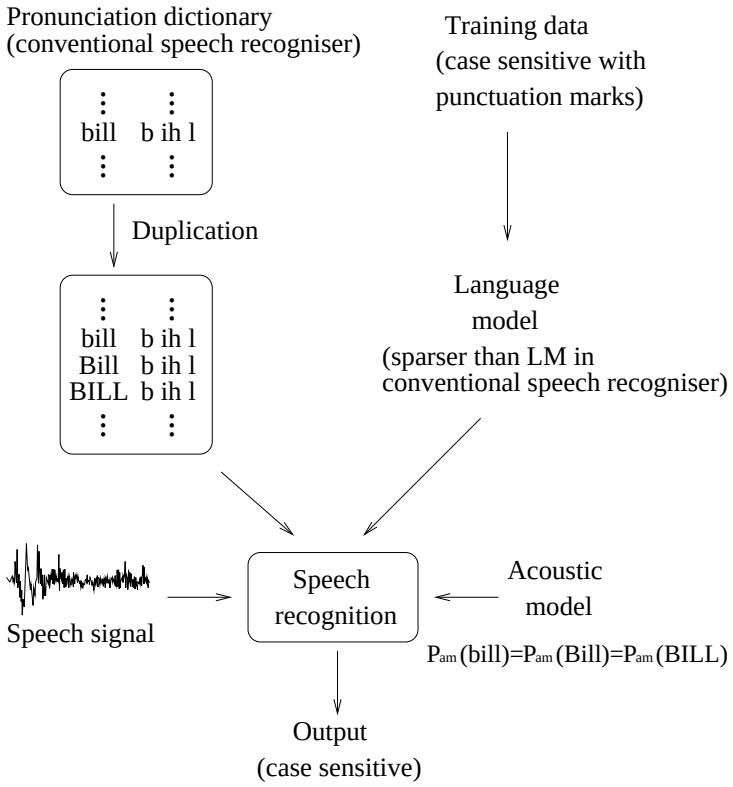

Figure 1: Overall procedures of the capitalisation generation system modified from speech recogniser.

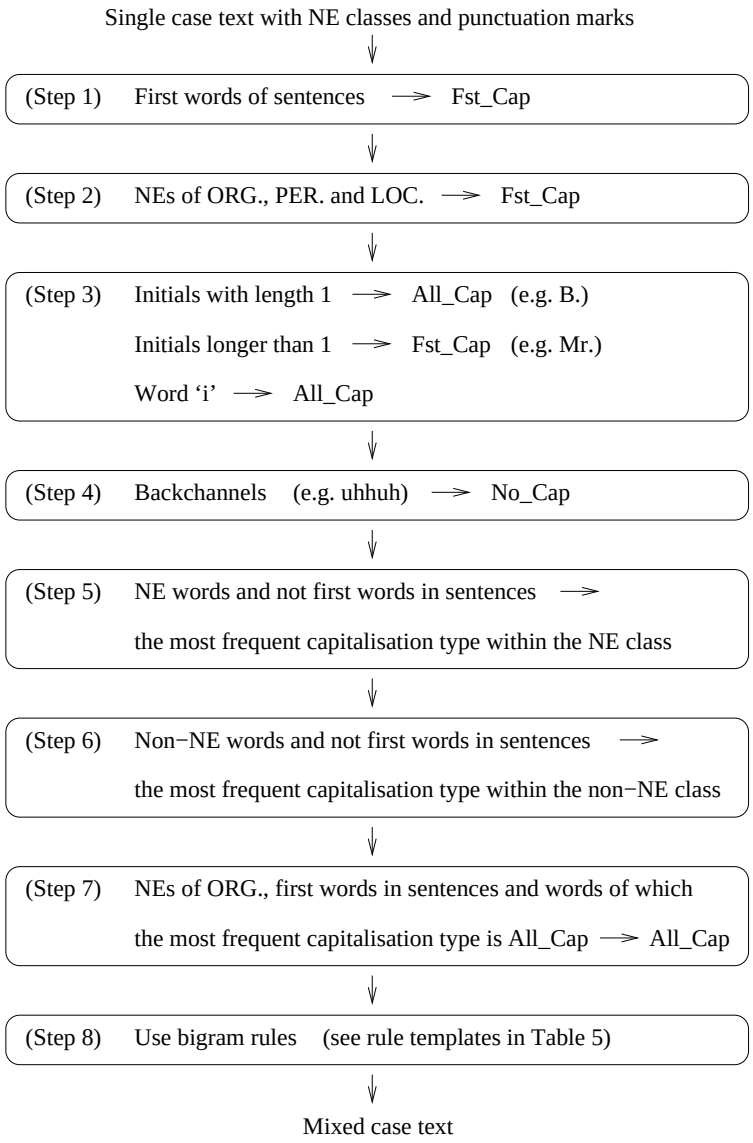

Figure 2: Procedures of the capitalisation generation system based on NE recognition and punctuation generation.

Name	Description	#Words	Purpose	Acoustic data
BNText92_97	1992_97 BN texts	184M	Training data	Not available
BNAcoustic98	100 hrs of Hub-4 data (1998)	774K	Training data	Available
TestBNAcoustic98	1998 benchmark test data	32K	Test data	Available

Table 1: Descriptions of the broadcast news training and test data.

Type	Description
No_Cap	Every character is de-capitalised
All_Cap	All characters are capitalised
Fst_Cap	Only first character is capitalised

Table 2: Categories of capitalisation types.

Word type			BNAcoustic98		TestBNAcoustic98	
NE class	Capitalisation type	Example	#FW	#non-FW	#FW	#non-FW
NE	No_Cap	far east	16	12,110	0	615
NE	All_Cap	B. B. C.	536	10,535	20	577
NE	Fst_Cap	Clinton	3,529	43,459	143	1,790
non-NE	No_Cap	sentence	1,587	638,477	24	26,134
non-NE	All_Cap	C. E. O.	2,842	6,887	83	141
non-NE	Fst_Cap	American	37,659	16,256	1,603	465

Table 3: Number of occurrences of each type of word based on the position of words in a sentence in BNAcoustic98 and TestBNAcoustic98. (FW: a first word in a sentence, non-FW: not a first word in a sentence)

	Number of occurrences	
Type	BNAcoustic98	TestBNAcoustic98
Words (any type)	773,893	31,595
Capitalised words	121,703	4,822
NE words	70,230	3,149
Single letter initial words (NE)	10,200	543
Single letter initial words (non-NE)	2,099	78
Sentences	46,169	1,873

Table 4: Statistics of BNAcoustic98 and TestBNAcoustic98.

Rule templates		
$w_0\ w_1,$	$w_0\ w_{-1},$	$w_0\ t_1$
$w_0\ t_{-1},$	$w_0\ c_1,$	$w_0\ c_{-1}$

Table 5: The rule templates used in bigram rule generation for capitalisation generation (w: words; t: NE types; c: capitalisation types). Subscripts define the distance from the current word.

System	WER	WER″	Precision	Recall	F-measure	SER
ModSR	22.97	17.27	0.7736	0.6942	0.7317	48.55

Table 6: Results of capitalisation generation for TestBNAcoustic98 using the system modified from the HTK system. (WER: Word Error Rate (%); WER″: WER after punctuation marks are removed and words in both reference and hypothesis are changed to single case; SER: Slot Error Rate (%))

System	Test condition			Result					
	Word	NE	Punc.	WER	WER″	Precision	Recall	F-measure	SER
NEPuncBased	Gen.	Gen.	Gen.	22.55	16.86	0.8094	0.6826	0.7406	45.93

Table 7: Results of the capitalisation generation system based on NE recognition and punctuation generation. (Punc.: Punctuation; Gen.: Generated; WER: Word Error Rate (%); WER″: WER after punctuation marks are removed and words in both reference and hypothesis are changed to single case; SER: Slot Error Rate (%))

System	Test condition			Result			
	Word	NE	Punc.	Precision	Recall	F-measure	SER(%)
NEPuncBased	Ref.	Ref.	Ref.	0.9726	0.9786	0.9756	4.89

Table 8: Results of the capitalisation generation system based on NE classes and punctuation marks for reference word sequences, NE classes and punctuation marks. (Punc.: Punctuation; Ref.: Reference)

Included step								Result			
1	2	3	4	5	6	7	8	Precision	Recall	F-measure	SER(%)
I								0.9818	0.3814	0.5494	62.57
I	I							0.8944	0.8004	0.8448	29.41
I	I	I						0.9581	0.8881	0.9218	15.08
I	I	I	I					0.9632	0.8881	0.9241	14.58
I	I	I	I	I				0.9817	0.9019	0.9401	11.45
I	I	I	I	I	I			0.9703	0.9681	0.9692	6.16
I	I	I	I	I	I	I		0.9705	0.9683	0.9694	6.12
I	I	I	I	I	I	I	I	0.9726	0.9786	0.9756	4.89

Table 9: Results of capitalisation generation system based on NE classes and punctuation marks with different combinations of processing steps for reference transcription.

Included step								Result			
1	2	3	4	5	6	7	8	Precision	Recall	F-measure	SER(%)
I								0.6256	0.1339	0.2206	93.74
I	I							0.7434	0.5308	0.6194	59.95
I	I	I						0.7913	0.5932	0.6781	54.45
I	I	I	I					0.7964	0.5932	0.6799	54.16
I	I	I	I	I				0.8057	0.6034	0.6901	52.49
I	I	I	I	I	I			0.8052	0.6738	0.7337	47.03
I	I	I	I	I	I	I		0.8054	0.6740	0.7339	46.99
I	I	I	I	I	I	I	I	0.8094	0.6826	0.7406	45.93

Table 10: Results of a capitalisation generation system based on NE classes and punctuation marks with different combinations of processing steps for speech recognition output when every capitalisation procedure is fully automated.

System	Test condition			Result			
	Word	NE	Punc.	Precision	Recall	F-measure	SER(%)
NEPuncBased	Ref.	Gen.	Ref.	0.9552	0.9643	0.9598	8.09

Table 11: Results of capitalisation generation system based on NE classes and punctuation marks for reference word sequences, generated NE classes and reference punctuation marks. (Punc.: Punctuation; Ref.: Reference; Gen.: Generated)

System	Test condition			Result			
	Word	NE	Punc.	Precision	Recall	F-measure	SER(%)
NEPuncBased	Ref.	Ref.	Gen.	0.8832	0.8861	0.8847	23.10

Table 12: Results of capitalisation generation system based on NE classes and punctuation marks for reference word sequences, reference NE classes and generated punctuation marks. (Punc.: Punctuation; Ref.: Reference; Gen.: Generated)

System	Test condition			Result			
	Word	NE	Punc.	Precision	Recall	F-measure	SER(%)
NEPuncBased	Ref.	Gen.	Gen.	0.8667	0.8715	0.8691	26.25

Table 13: Results of capitalisation generation system based on NE classes and punctuation marks for reference word sequences, generated NE classes, and generated punctuation marks. (Punc.: Punctuation; Ref.: Reference; Gen.: Generated)

System	Test condition			Result			
	Word	NE	Punc.	Precision	Recall	F-measure	SER(%)
NEPuncBased	Ref.	Gen.	Ref.	0.9588	0.9608	0.9598	8.04
MS Word 2000	Ref.	N/A	Ref.	0.9987	0.8045	0.8911	19.66

Table 14: Results of capitalisation generation by NEPuncBased for reference word sequences, generated NE classes and reference punctuation marks using 10.7% of TestBNAcoustic98. These results are compared with those from Microsoft Word for the same part of TestBNAcoustic98. (Punc.: Punctuation; Ref.: Reference; Gen.: Generated)

References

D. Beeferman, A. Berger & J. Lafferty (1998). Cyberpunc: A Lightweight Punctuation Annotation System for Speech. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 689–692, 1998.

D. Bikel, S. Miller & R. Schwartz (1997). Nymble: a High-Performance Learning Namefinder. In *Proceedings of the Applied Natural Language Processing*, pages 194–201, 1997.

L. Breiman, J. H. Friedman, R. A. Olshen & C. J. Stone (1983). *Classification and Regression Trees*. Wadsworth and Brooks, 1983.

E. Brill (1993). *A Corpus-Based Approach to Language Learning*. PhD thesis, University of Pennsylvania, 1993.

E. Brill (1994). Some Advances in Rule-Based Part of Speech Tagging. In *Proceedings of the 12th National Conference on Artificial Intelligence*, pages 722–727, 1994.

C. Chen (1999). Speech Recognition with Automatic Punctuation. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 447–450, 1999.

H. Christensen, Y. Gotoh & S. Renals (2001). Punctuation Annotation using Statistical Prosody Models. In *Proceedings of the ISCA Workshop on Prosody in Speech Recognition and Understanding*, 2001.

Y. Gotoh, S. Renals & G. Williams (1999). Named Entity Tagged Language Models. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 513–516, 1999.

Y. Gotoh & S. Renals (2000). Sentence Boundary Detection in Broadcast Speech Transcripts. In *Proceedings of the International Workshop on Automatic Speech Recognition*, pages 228–235, 2000.

R. Grishman & B. Sundheim (1995). Design of the MUC-6 Evaluation. In *Proceedings of the 6th Message Understanding Conference*, pages 1–11, 1995.

J. Huang & G. Zweig (2002). Maximum Entropy Model for Punctuation Annotation from Speech. In *Proceedings of the International Conference on Spoken Language Processing*, 2002.

J. Kim & P. C. Woodland (2000a). Rule Based Named Entity Recognition. Technical Report CUED/F-INFENG/TR.385, Cambridge University Engineering Department, 2000.

J. Kim & P. C. Woodland (2000b). A Rule-based Named Entity Recognition System for Speech Input. In *Proceedings of the International Conference on Spoken Language Processing*, pages 521–524, 2000.

J. Kim & P. C. Woodland (2001). The Use of Prosody in a Combined System for Punctuation Generation and Speech Recognition. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 2757–2760, 2001.

F. Kubala, R. Schwartz, R. Stone & R. Weischedel (1998). Named Entity Extraction from Speech. In *Proceedings of the Broadcast News Transcription and Understanding Workshop*, pages 287–292, 1998.

C. J. Leggetter & P. C. Woodland (1995). Maximum Likelihood Linear Regression for Speaker Adaptation of Continuous Density Hidden Markov Models. *Computer Speech and Language*, 9:171–185, 1995.

J. Makhoul, F. Kubala, R. Schwartz & R. Weischedel (1999). Performance Measures for Information Extraction. In *Proceedings of the DARPA Broadcast News Workshop*, pages 249–252, 1999.

A. Mikheev (1999). A Knowledge-free Method for Capitalized Word Disambiguation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics*, pages 159–166, 1999.

S. Miller, M. Crystal, H. Fox, L. Ramshaw & R. Schwartz (1997). Algorithms that Learn to Extract Information. BBN: Description of the SIFT System as Used for MUC-7. In *Proceedings of the 7th Message Understanding Conference*, 1997. Available at http://www.muc.saic.com/proceedings/muc_7_toc.html.

MUC (1995). Named Entity Task Definition. In *Proceedings of the 6th Message Understanding Conference*, pages 317–332, 1995.

NIST (1998a) 1998 NIST Hub-4 Information Extraction (Named Entity) Broadcast News Benchmark Test Evaluation. Available at ftp://jaguar.ncsl.nist.gov/csr98/h4iene_98_official_scores_990107/index.htm.

NIST (1998b) NIST Hub-4 IE scoring pipeline package version 0.7. Available at ftp://jaguar.ncsl.nist.gov/csr98/official-IE-98_scoring.tar.Z.

T. Niesler, E. Whittaker & P. C. Woodland (1998). Comparison of Part-Of-Speech and Automatically Derived Category-Based Language Models for Speech Recognition. In *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 177–180, 1998.

J. Odell, P. C. Woodland & T. Hain (1999). The CUHTK-Entropic 10xRT Broadcast News Transcription System. In *Proceedings of the DARPA Broadcast News Workshop*, pages 271–275, 1999.

D. Pallett, J. Fiscus, J. Garofolo, A. Martin & M. Przybocki (1999). 1998 Broadcast News Benchmark Test Results: English and Non-English Word Error Rate Performance Measures. In *Proceedings of the DARPA Broadcast News Workshop*, pages 5–12, 1999.

M. Przybocki, J. Fiscus, J. Garofolo & D. Pallett (1999). 1998 Hub-4 Information Extraction Evaluation. In *Proceedings of the DARPA Broadcast News Workshop*, pages 13–18, 1999.

S. Rayson, D. Hachamovitch, A. Kwatinetz & S. Hirsch (1998). Autocorrecting Text Typed into a Word Processing Document. 1998. U.S. patent 5761689. Available at http://www.delphion.com.

A. Stolcke & E. Shriberg (1996). Automatic Linguistic Segmentation of Conversational Speech. In *Proceedings of the International Conference on Spoken Language Processing*, pages 1005–1008, 1996.

A. Stolcke, E. Shriberg, R. Bates, M. Ostendorf, D. Hakkani, M. Plauche, G. Tür & Y. Lu (1998). Automatic Detection of Sentence Boundaries and Disfluencies based on Recognized Words. In *Proceedings of the International Conference on Spoken Language Processing*, pages 2247–2250, 1998.

V. Warnke, R. Kompe, H. Niemann & E. Nöth (1997). Integrated Dialog Act Segmentation and Classification using Prosodic Features and Language Models. In *Proceedings of the European Conference on Speech Communication and Technology*, pages 207–210, 1997.

P. C. Woodland, T. Hain, S. Johnson, T. Niesler, E. Whittaker & S. Young (1998). The 1997 HTK Broadcast News Transcription System. In *Proceedings of the Broadcast News Transcription and Understanding Workshop*, 1998.

P. C. Woodland, T. Hain, G. Moore, T. Niesler, D. Povey, A. Tuerk & E. Whittaker (1999). The 1998 HTK Broadcast News Transcription System: Development and Results. In *Proceedings of the DARPA Broadcast News Workshop*, pages 265–270, 1999.

P. C. Woodland (2002). The development of the HTK Broadcast News transcription system: An overview. *Speech Communication*, 37(1-2):47–67, 2002.