IMPLEMENTATION OF AUTOMATIC CAPITALISATION GENERATION SYSTEMS FOR SPEECH INPUT

Ji-Hwan Kim and P.C. Woodland

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

ABSTRACT

In this paper, two 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 first words in sentences or NE words. Both systems are compared for speech input. The system based on NE recognition and punctuation generation shows better results in Word Error Rate (WER) and in F-measure than the system modified from the speech recogniser.

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. When speakers are not aware that their speech is to be automatically transcribed, e.g. Broadcast News (BN), the system can not rely on the speaker to say “capitalise the current word” whenever necessary. The readability of speech recognition output would be greatly enhanced by generating proper capitalisation.

For text input, many commercial implementations of automatic capitalisation are provided with word processors. 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 [1]. In this implementation, whether the current word is at the start of a sentence was determined by a sentence capitalisation state machine. Capitalisation of words which are not the first words in sentences could be performed by dictionary look-up.

An approach to the disambiguation of capitalised words was presented in [2]. 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 tasks of Named Entity (NE) recognition, punctuation generation, and capitalisation generation are strongly related to each other, since most capitalised words are first words in sentences or NEs. The importance of NE recognition in automatic capitalisation was mentioned in [3]. NE recognition experiments showed that the performance deteriorates when the capitalisation and punctuation information are missing [4].

NE recognition systems are generally categorised as either stochastic (typically HMM-based) or rule-based. In [5], an automatic rule inference method is presented for the NE task. 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.

An automatic punctuation generation method consisting of a modified speech recogniser was proposed for BN data in [6]. In that paper, several straightforward modifications of a conventional speech recogniser allow the system to produce punctuation and speech recognition hypothesis simultaneously. Improvements were obtained by re-scoring multiple punctuated hypotheses using prosodic information.

In Section 2, we present two methods for automatic capitalisation generation. The experimental setup to evaluate capitalisation on BN data is described in Section 3. In Sections 3.1 and 3.2, experimental results are presented and discussed.

2. CAPITALISATION GENERATION

In this paper, two different automatic capitalisation generation systems are presented. The first system (S_SR) is a slightly modified speech recogniser. The other system (S_NE_P) is based on NE recognition and punctuation generation, since most capitalised words are first words in sentences or NE words.

These systems examine three types of capitalisation: all characters of a word are capitalised (All_Cap), only the first character of a word is capitalised (Fst_Cap), and every character of a word is de-capitalised (No_Cap). There is a small number of exceptional cases which are not categorised as any of these categories. Most of these are surnames. For example, McWethy, MacLaine, O’Brien, LeBowe and JonBenet. These exceptional cases are classified as Fst_Cap.

2.1. System modified from speech recogniser (S_SR)

Three small modifications to a conventional speech recognition system are required to produce case sensitive outputs:

1. Every word in its vocabulary is duplicated for the three different capitalisation types (All_Cap, Fst_Cap, No_Cap).
2. Every word in its pronunciation dictionary is duplicated with its pronunciation in the same way as used for the vocabulary duplication.
3. The language model is re-trained on mixed case texts.

Figure 1 illustrates the overall capitalisation generation system, modified from a conventional speech recogniser. The same

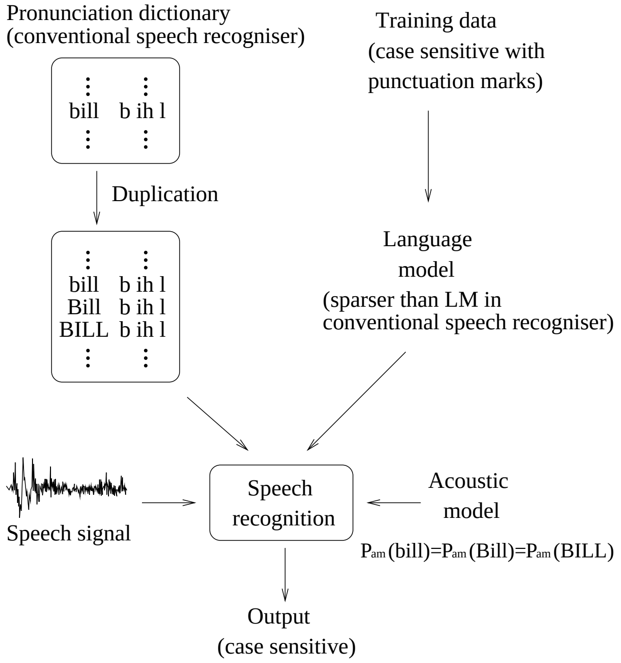

Fig. 1. Overall procedures of the capitalisation generation system modified from speech recogniser

acoustic score will be assigned to duplicated words, since they have the same pronunciations. However, different hypotheses will be generated using the different language model scores.

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 language model is trained on mixed-case texts which contain punctuation marks.

The correlation between punctuation and pauses was investigated in [7]. These experiments showed that pauses closely correspond to punctuations. The correlation between pause lengths and sentence boundary marks was studied for broadcast news data in [8]. In that study, it was observed that the longer the pause duration, the greater the chance of a sentence boundary existing. 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 [6].

2.2. System based on NE recognition and punctuation generation (S_NE_P)

The method of capitalisation generation presented in this section is based on NE recognition and punctuation generation, since most capitalised words are the first words in a sentence or NE words. S_NE_P uses the rule-based (transformation-based) NE recognition system [5], which uses the Brill rule [9] inference approach, and the punctuation generation system which incorporates prosodic information along with acoustic and language model information [6].

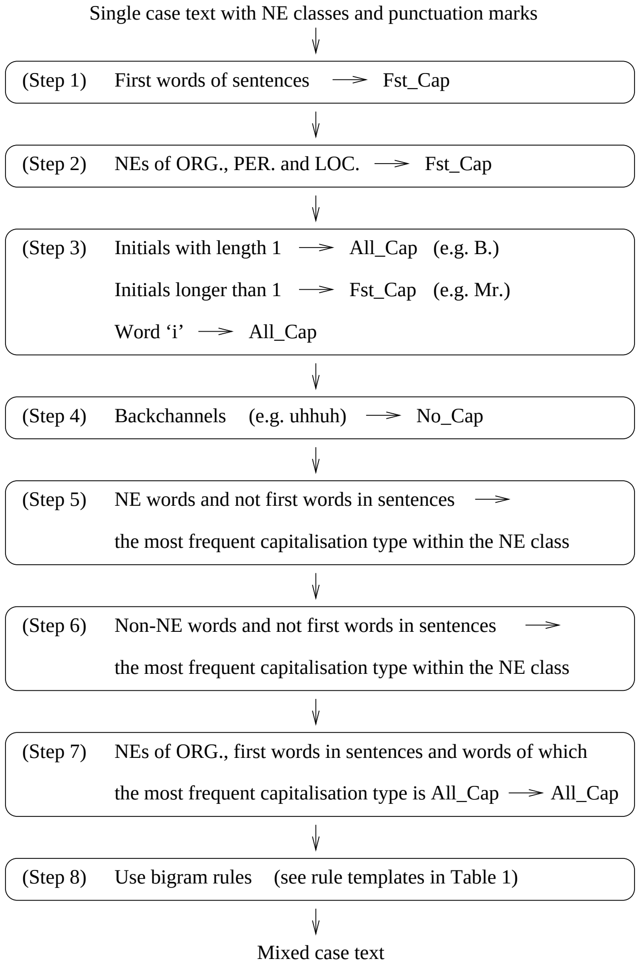
Single case text with NE classes and punctuation marks
(Step 1) First words of sentences → Fst_Cap
(Step 2) NEs of ORG., PER. and LOC. → Fst_Cap
(Step 3) Initials with length 1 → All_Cap (e.g. B.)
Initials longer than 1 → Fst_Cap (e.g. Mr.)
Word 'i' → All_Cap
(Step 4) Backchannels (e.g. uhhuh) → No_Cap
(Step 5) NE words and not first words in sentences → the most frequent capitalisation type within the NE class
(Step 6) Non-NE words and not first words in sentences → the most frequent capitalisation type within the NE class
(Step 7) NEs of ORG., first words in sentences and words of which the most frequent capitalisation type is All_Cap → All_Cap
(Step 8) Use bigram rules (see rule templates in Table 1)
Mixed case text

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

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. 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.

Punctuation generation uses two straightforward modifications of a conventional speech recogniser described in Section 2.1. 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

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 1. 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

using a prosodic feature model based on Classification And Regression Trees (CART) [10].

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 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 ideas used in the development of the rule-based NE recognition system in [5] are applied in the automatic generation of these rules for capitalisation generation.

Six rule templates are used for the generation of bigram rules for capitalisation generation. These six rule templates are shown in Table 1. 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.

Particular importance must be given to the effect of words encountered in the test data which have not been seen in the training data. One way of improving 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 1 are applied one-by-one in step 8 according to a given order.

Name	Description	#Words
DB92_97	1992_97 BN texts	184M
DB98	100 hrs of Hub-4 data (1998)	774K
TDB98	1998 benchmark test data	32K

Table 2. Database descriptions

3. EXPERIMENTS

Broadcast News data provides a good test-bed for speech recognition, because it requires systems to handle unknown speakers, a large vocabulary, and various domains. In this paper, we use a 184 million words of BN text from the period of 1992-1997 inclusive[1] and a 71-hour BN acoustic training data set (acoustic data and its transcription) released for the 1998 Hub-4 evaluation as training data. We also used 3 hours of test data from the NIST 1998 Hub-4 BN benchmark tests. Table 2 summarises the training and test data. 4-gram language models were trained by interpolating language models trained on DB92_97 and DB98 using a perplexity minimisation method.

The systems were evaluated using the agreement between the capitalisation types in the hypothesis file and those in the reference file. Precision and Recall were used as metrics for assessing the performance. These are defined as:

$$P = \frac{\text{number of correct capitalised words}}{\text{number of hypothesised capitalised words}} \tag{1}$$

and

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

A half score is given when a word is capitalised, but generated as a different type of capitalisation. The F-measure is the uniformly weighted harmonic mean of Precision and Recall:

$$F = \frac{PR}{(P+R)/2} \tag{3}$$

The F-measure is also used as a metric for assessing performance. To implement scoring, the version 0.7 of NIST HUB-4 scoring pipeline [11] 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.

3.1. Results: S_SR

The first automatic capitalisation system was implemented by small modifications to the HTK Broadcast News (BN) transcription system. Details of the HTK BN transcription system are given in [12].

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 DB92_97 and DB98.

Table 3 shows the results of capitalisation generation for TDB98 using this system. When Word Error Rate (WER) is measured, words are changed into single case from reference and hypothesis in order to measure the pure speech recognition rate. As

[1] The 1992-1996 part was provided by the LDC and the 1997 part by Primary Source Media.

System	WER(%)	WER$''$(%)	P	R	F
S_SR	22.97	17.27	0.7736	0.6942	0.7317
S_NE_P	22.55	16.86	0.8094	0.6826	0.7406

Table 3. Results of capitalisation generation for TDB98 using S_SR and S_NE_P. (WER$''$: WER after punctuation is removed. WER and WER$''$ are measured on single case reference and hypothesis)

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.

For punctuation generation, the HTK system gave 22.73% of WER in [6]. The difference between WER in punctuation generation and that in capitalisation generation is measured as 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. Distortion of LM: In many cases, the first word of a sentence is not a NE. Most of these words are not capitalised, if they are used in the middle of sentences.
2. Sparser LM: As the size of vocabulary is increased, LMs are more sparse and estimating probabilities of word sequences becomes more difficult. Also, the search space is increased.

3.2. Results: S_NE_P

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. Punctuation marks are generated first by the punctuation generation system in [6], which incorporates prosodic information along with acoustic and language model information. Then NE recognition is performed using the rule-based NE recognition system in [5], which generates rules automatically.

The automatic punctuation generation system used in this capitalisation generation system gives an F-measure of 0.4239 for punctuation marks when WER is minimised [6]. At this point, this punctuation generation system gave a WER of 22.55%. The rule-based NE recogniser trained under the condition of ‘with punctuation and name lists but without capitalisation’ was used. This NE recogniser reported an F-measure of 0.9007 for the reference transcription of TDB98 [5].

The frequency table and bigram rules were constructed using the transcription of DB98, because DB98 is the only training data which is provided with reference NE classes. Table 3 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 and WER$''$. Compared to the other capitalisation generation system (S_SR), this system (S_NE_P) shows better results by: 0.42% in WER, 0.41% in WER$''$, and 0.0089 in F-measure. The factors which cause these differences were explained as ‘the distortion of LM’ and ‘sparser LM’ in Section 3.1.

The contribution of each experimental step was measured. By just performing step 2, in addition to step 1, an F-measure of 0.6194 is already obtained, although the recall (0.5308) is poor compared to the precision (0.7434). Adding steps 3 and 4, which can be done by straightforward processes without the need for training data, an F-measure of 0.6799 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 0.7339. This is increased to 0.7406 points in F-measure using bigram rules.

4. CONCLUSIONS

In this paper, two different systems have been proposed for the task of automatic capitalisation generation. The first is a slightly modified speech recogniser. The other system is based on NE recognition and punctuation generation. In order to compare the performance of the proposed systems, experiments of automatic capitalisation generation were performed for speech input. The system based on NE recognition and punctuation generation showed better results in WER and in F-measure than the system modified from the speech recogniser, because the latter system has a distorted LM and a sparser LM.

5. ACKNOWLEDGEMENTS

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

6. REFERENCES

[1] S. Rayson, D. Hachamovitch, A. Kwatinetz, and S. Hirsch, “Autocorrecting Text Typed into a Word Processing Document,” 1998, U.S. patent 5761689. Available at http://www.delphion.com.

[2] A. Mikheev, “A Knowledge-free Method for Capitalized Word Disambiguation,” in *Proc. Annual Meeting of the Association for Computational Linguistics*, 1999, pp. 159–166.

[3] Y. Gotoh, S. Renals, and G. Williams, “Named Entity Tagged Language Models,” in *Proc. ICASSP*, 1999, pp. 513–516.

[4] F. Kubala, R. Schwartz, R. Stone, and R. Weischedel, “Named Entity Extraction from Speech,” in *Proc. Broadcast News Transcription and Understanding Workshop*, 1998, pp. 287–292.

[5] J. Kim and P. C. Woodland, “A Rule-based Named Entity Recognition System for Speech Input,” in *Proc. ICSLP*, 2000, pp. 521–524.

[6] J. Kim and P. C. Woodland, “The Use of Prosody in a Combined System for Punctuation Generation and Speech Recognition,” in *Proc. Eurospeech*, 2001.

[7] C. Chen, “Speech Recognition with Automatic Punctuation,” in *Proc. Eurospeech*, 1999, pp. 447–450.

[8] Y. Gotoh and S. Renals, “Sentence Boundary Detection in Broadcast Speech Transcripts,” in *Proc. International Workshop on Automatic Speech Recognition*, 2000, pp. 228–235.

[9] E. Brill, *A Corpus-Based Approach to Language Learning*, Ph.D. thesis, University of Pennsylvania, 1993.

[10] L. Breiman, J. H. Friedman, R. A. Olshen, and C. J. Stone, *Classification and Regression Trees*, Wadsworth and Brooks, 1983.

[11] “NIST Hub-4 IE scoring pipeline package version 0.7,” Available at ftp://jaguar.ncsl.nist.gov/csr98/official-IE-98_scoring.tar.Z.

[12] P. C. Woodland, T. Hain, G. Moore, T. Niesler, D. Povey, A. Tuerk, and E. Whittaker, “The 1998 HTK Broadcast News Transcription System: Development and Results,” in *Proc. DARPA Broadcast News Workshop*, 1999, pp. 265–270.