Restoration of Case Information in All-Cap English Broadcast Transcription

Yu-Ting Liang
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, 300, Taiwan, ROC
u902518@cs.nthu.edu.tw

Jian-Chen Wu
Department of Computer Science
National Tsing Hua University
101, Kuangfu Road, Hsinchu, 300, Taiwan, ROC
g904374@cs.nthu.edu.tw

The local broadcast ICRT (International Community Radio Taipei) in Taipei has their news scripts sent to their listeners in ALL CAPS, which makes the articles more difficult to read. Therefore, we think it may facilitate the readers if we transform the text into normal cases that we are familiar with. In this prototype system, we established a practical method of restoration of case information, using different techniques from NLP and statistics. The system can apply many different kinds of approach, however, in this prototype, we focus our analysis and test data on broadcast transcription.

Basically, our research involves:

- Establishing a very large database containing numerous vocabularies and uses as our training data.
- Obtaining text from ICRT news scripts sent by e-mails as our test data.
- Restoring the cases of the contents into cases that we are more acquainted with.
- Handling some exceptions.

Establishing a very large database. Our training data comes from VOA news, which consists of 9138 articles, 3 million words in total. For each article, we segment its contents into individual words and calculate their n-gram probabilities.

Obtaining text from ICRT news scripts. We perform similar piecing process on the news scripts. After obtaining each isolated word, we query its probabilities in unigram, bigram, and trigram probabilities, which have two, four, and eight values respectively from our training data.

Restoring the cases of the contents. After accomplishing Viterbi algorithm (Rabiner, 1989) to compute the highest probability and its P-model value (Lucian Vlad Lita, 2003), we acquire the best restoration of case for each word, and then we alter the texts. We have an example in Figure 1, an original text from one of the ICRT news scripts and the text after restoration.

Upper Case: WILLIAM SAMPSON SPENT 31 MONTHS IN PRISON IN SAUDI ARABIA, WHERE HE WAS SENTENCED TO DEATH OVER A SERIES OF BOMBINGS THAT KILLED ONE PERSON.

After Restoration: William Sampson spent 31 months in prison in Saudi Arabia, where he was sentenced to death over a series of bombings that killed one person.

Figure 1: An example of upper case text and restoration

Handling exceptions. Actually, the word 'Sampson' was not found in our training data, however, we assume unknown words as proper nouns and therefore we capitalize its first letter. Here we have another experiment in Figure 2.

Upper Case: THE U.S. MILITARY ISSUED A PUBLIC APOLOGY TO THE PEOPLE OF A SHIITE MUSLIM NEIGHBORHOOD IN BAGHDAD ON THURSDAY FOR AN INCIDENT IN WHICH A MAN WAS KILLED AND FOUR OTHERS WOUNDED AFTER AN AMERICAN BLACK HAWK HELICOPTER BLEW DOWN AN ISLAMIC BANNER WITH ITS ROTOR WASH. THAT APPEARS TO BE A MAJOR SHIFT IN THE MILITARY'S RELATIONS WITH THE NEWS MEDIA.

After Restoration: **the** U. S. military issued a public apology to the people of a Shiite Muslim neighborhood in Baghdad on Thursday for an incident in which a man was killed and four others wounded after an American Black Hawk helicopter blew down an Islamic banner with its rotor wash. **that** appears to be a major shift in the military's relations with the news media.

Figure 2: Another example of upper case text and restoration

Again, we found some adjustments have to be done, and the first letter of the first word in a sentence ought to be in upper case is one of them. Even so, we have to ask ourselves, "What is a sentence?" Is it something ends up with a period, an exclamation mark, or a question mark? Apparently, we can find a counter example with "U.S.". Here we use heuristic sentence boundary detection algorithm to determine what a sentence is and capitalized the first words in a

sentence as shown in Figure 3.

After restoration: **The** U.S. military issued a public apology to the people of a Shiite Muslim neighborhood in Baghdad on Thursday for an incident in which a man was killed and four others wounded after an American Black Hawk helicopter blew down an Islamic banner with its rotor wash. **That** appears to be a major shift in the military's relations with the news media.

Figure 3: The previous example after sentence adjustment

Our demonstration model shows we can convert all-cap English news scripts quite well. There are some possible improvements and our future works are improving our performance, which can reduce the time we spend on transforming the text. Also, create a macro in Outlook so if the readers receive their e-mails from ICRT with Outlook, they may have the restoration done by running a macro. We are looking forward to finding the readers feeling this tool useful and somewhat convenient.

Acknowledgements

We acknowledge the support of NSC under contract number: 92-2815-C-007 -004 -E . Many thanks are due to Dr. Jason S. Chang for his guidance in NLP and ICRT for their news scripts.

References

Lucian Vlad Lita, 2003. *tRuEcasIng.*

Hai Leong Chieu, Hwee Tou Ng, 2002 Teaching a Weaker Classifier: Named Entity Recognition on Upper Case Text.

Andrei Mikheev, 1999 A Knowledge-free Method for Capitalized Word Disambiguation.

Alison Huettner, Pero Subasic, 2000 Fuzzy Typing for Document Management.

Christopher D. Manning and Hinrich Schutze. Foundations of statistical natural language processing, 2000, pp. 123-136