A Semi-automatic Approach to Identifying and Unifying Ambiguously Encoded Arabic-Based Characters

Sardar Jaf
sardar.jaf@durham.ac.uk
School of Engineering & Computer Sciences, Durham University

Motivation

One of the main resources available for processing resource-scarce languages is raw text collected from the Internet. Mandy of those languages use a modified version of Arabic characters, where some of them have the same form but with different Unicode values (ambiguous characters), which leads to word duplication, hence it is important to identify and unify ambiguous characters in the normalisation stage.

The Challenges of Processing Kurdish

1. There are many dialects in Kurdish. Some dialects use A modified version of Arabic Alphabet. [1]
2. A letter, such as { ه } constitutes H, h, E, or e depending on its location in the word. If it appears in position four in the word its Unicode value is u06BE, but in some cases it is assigned u0647. If it appears in position five and nine, its Unicode value is either u0647, u06BE, or u06D5.
3. Inconsistent encoding leads to large numbers of words losing their unique form. Table I. contains examples of Kurdish characters with different Unicode values.

Words	Frequency	Unicode value
ه /ha/, /a/, zero-width non-joiner character	2361391	u06D5
	1961352	u0647
	51442	u06BE
ی /ye/	2481987	u06CC
	69363	u0649
ک /k/	585728	u06A9
	537621	u0643

Table I. Words with ambiguous character

Dataset

21,000 news articles were collected from online news websites. This amounted to about 2,000,000 words. Words that have occurred less than 0.0001% in the data were removed. These words were either non-Kurdish words; words with incorrect spelling; or words that are accidentally merged with some other words during the parsing process of the web pages.

Identifying Unicode Characters

- Extract all the unique words in the dataset
- Save all the unique words in a second lexicon
- Process the characters of the recorded unique words. Identify the unique characters with their Unicode values
- This results in 42,987 words and the processing time was 26 seconds.

The unification process

We generated a mapping dictionary that mapped the Unicode value of ambiguous characters to different Unicode value, as shown in Table II.

Unicode Value	Mapped Unicode value
u0647	u06BE or u06D5
u0649	u06CC
u0643	u06A9

Table II. Mapping one Unicode to another

Evaluation

All the unique characters and their Unicode values from the processed lexicon extracted. The characters were manually inspecting. The absence of ambiguous characters indicated that all characters in the lexicon were encoded correctly.

Conclusions

Our study involved identifying Ambiguous characters and using a mapping document to unify them. We have applied our approach primarily to Kurdish and Farsi. In the future, we plan to apply this approach to other related languages such as Urdu and Pashtu.

[1] D. N. MacKenzie, Kurdish dialect studies, volume 1 of London Oriental Series. Oxford University Press, 1961.