Nr.: FIN-001-2017

Manually Annotated Spelling Error Corpus for Amharic

Andargachew Mekonnen Gezmu, Binyam Ephrem Seyoum, Tirufat Tesifaye Lema, Andreas Nürnberger

Data and Knowledge Engineering Group
Nr.: FIN-001-2017

Manually Annotated Spelling Error Corpus for Amharic

Andargachew Mekonnen Gezmu, Binyam Ephrem Seyoum, Tirufat Tesfaye Lema, Andreas Nürnberger

Data and Knowledge Engineering Group
Impressum (§ 5 TMG)

Herausgeber:
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Der Dekan

Verantwortlich für diese Ausgabe:
Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Andargachew Mekonnen Gezmu
Postfach 4120
39016 Magdeburg
E-Mail: andargachew.gezmu@ovgu.de

http://www.cs.uni-magdeburg.de/Technical_reports.html
Technical report (Internet)
ISSN 1869-5078

Redaktionsschluss: 23.10.2017

Bezug: Otto-von-Guericke-Universität Magdeburg
Fakultät für Informatik
Dekanat
Manually Annotated Spelling Error Corpus for Amharic

Andargachew Mekonnen Gezmu
Otto-von-Guericke-Universität
Magdeburg, Germany
andargachew.gezmu@ovgu.de

Binyam Ephrem Seyoun
Addis Ababa University
Addis Ababa, Ethiopia
binyam.ephrem@aaau.et

Tirufat Tesifaye Lema
Hawassa University
Hawassa, Ethiopia
tiruffat@hu.edu.et

Andreas Nürnberger
Otto-von-Guericke-Universität
Magdeburg, Germany
andreas.nuernberger@ovgu.de

Abstract—This paper presents a manually annotated spelling error corpus for Amharic, lingua franca in Ethiopia. The corpus is designed to be used for evaluation of spelling error detection and correction. The misspellings are tagged as non-word and real-word errors. In addition, the contextual information available in the corpus makes it useful in dealing with both types of spelling errors.

Keywords—spelling error, corpus, Amharic

I. INTRODUCTION

Amharic, the official working language of the Federal Government of Ethiopia, is the second-most widely spoken Semitic language next to Arabic. It is also spoken by hundreds of thousands of Ethiopian immigrants in North America and Ethiopian Jews in Israel [1] [2]. In spite of its great number of speakers and being a written language for more than five centuries, Amharic is still a scarce resource language as far as computational resources are concerned.

For researchers who are concerned about the development of Amharic spelling error correctors, there is no standard error corpus to evaluate their approach. Spelling error corpora can be collected automatically from keystroke logs or word-typing games. Baba and Suzuki [3] extracted pairs of misspellings and corrections from input logs by using Amazon’s Mechanical Turk. Researchers also attempt to collect such corpora from word-typing games [4] [5]. However, the former study is limited to languages supported by the crowdsourcing, and in the latter studies, subjects may behave differently than writing a regular text. Thus, we have developed a manually annotated spelling error corpus for Amharic.

II. RELATED WORK

Grudin [6] offers a detailed study of the general pattern of errors made by novice and expert typists, where the subjects are asked to transcribe a text without correcting the errors they made. The author compiled letter confusion matrices in which typographical errors are categorized according to the letter intended and the letter actually struck. Even though it might be used in analyzing and modeling sources of misspellings, its lack of contextual information limits its scope of usage particularly for real-word errors. A manually tagged spelling error corpus with contextual information is available from the book “English for the Rejected” [7] in the Oxford Text Archive [8]. Despite it was originally handwritten by poor spellers, its contextual information makes it still useful for evaluation purposes. Furthermore, Mitton [9] argues that misspellings produced by poor spellers are worse than those caused by typing slips; and a spelling corrector that can deal with poor spellings has a good chance of handling typos, but the reverse might not be true.

III. TYPES OF SPELLING ERRORS

The spelling errors in Amharic can be grouped as non-word and real-word errors. When typographical or cognitive errors accidently produce valid Amharic words we get real-word errors, otherwise, we get non-word errors. Typographical errors include insertion, deletion, transposition, substitution of letters. Missed out spaces are also sources of typos.

The cognitive errors in Amharic mainly result from the inconsistency of its writing system, Ethiopic. Though Ethiopic shares most feature of abugida, it is considered as syllabary [10] [11] [12]. Amharic has 27 consonant phonemes and seven vowels. Four of these phonemes have one or more homophonic character representations. The homophonic characters are the source of many cognates —some scholars consider them as homonyms— (e.g., ክባት [ba], ካጥ [sa], ካሃ and ካ[a] not with their variants [13]. As such, real-word errors might occur from wrongly typed homonyms. For example, ኮወል [s’i’i] meaning “paint” as its origin is ꗍ{sɨ’ɨl} “sun”). The general practice for strict Amharic writing style is that spellings of Amharic words inherited from Ge’ez, a parent language of Amharic, should follow Ge’ez features as much as possible, and loan words that use homophonic characters should be written only with ሥ [ba], ኪ [sa], ጠ [s ‘i’] and ከ [a] not with their variants [13]. As such, real-word errors might occur from wrongly typed homonyms. For example, ኮወል [s’i’i] is a real-word error for ኮወል ኮስል [s’i’i] meaning “paint” as its origin is the Ge’ez word ኮስል. However, in modern Amharic writings such as newspapers and magazines, the homophonic characters are commonly observed to be used interchangeably. Since our source data are from such sources we designed our guidelines in a way to reflect the misspellings and their intended spellings.

IV. GUIDELINES

We set guidelines to properly annotate misspellings collected from different sources with their contextual information. The guidelines are as follows:

- If a misspelling is not a valid Amharic word, tag it as a non-word error;
- If a valid Amharic word is determined to be a misspelling based on its neighboring words context, tag it as a real-word error;
- When deriving corrections of misspellings, adhere to intended spellings of the original authors rather than the strict Amharic writing style;
- Tag all words that result from informal Amharic dialects as non-word errors; and

1 We used the International Phonetic Alphabet (IPA) for Amharic pronunciation [17] [18]
Follow this format to tag misspellings: `<ERR target=word type=type> misspelling </ERR>`. Where `target` is the correction of the misspelling and `type` is the type for the misspelling, either non-word or real-word.

V. DATA SOURCES

The data sources are textual documents obtained from random samples of Amharic news articles of Deutsche Welle and Voice of America in the period of June 6 up to November 29, 2016; a retyped document of Aklilu [14]; and errata list of an Amharic novel by Alemayehu [15]. Totally 367 sentences are annotated with guidelines presented in the previous section. The annotated corpus constitutes the appendix of this paper.

VI. RESULTS

This section presents the number of misspellings by their types, the Damerau-Levenshtein edit distance [16] of the misspellings from their corrections, and the correlation between the misspellings and their string lengths.

Among the 372 misspellings, 287 (77.15%) are found to be non-word and 85 (22.85%) are real-word spelling errors. Two of the real-word and 34 of the non-word misspellings occur twice in the documents.

TABLE I. THE DAMERAU-LEVENSHTEIN EDIT DISTANCE OF THE MISSPELLINGS AND THEIR CORRECTIONS

<table>
<thead>
<tr>
<th>Edit Distance</th>
<th>Count</th>
<th>Percentage</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>290</td>
<td>77.96%</td>
</tr>
<tr>
<td>2</td>
<td>59</td>
<td>15.86%</td>
</tr>
<tr>
<td>3</td>
<td>18</td>
<td>4.84%</td>
</tr>
<tr>
<td>4</td>
<td>5</td>
<td>1.34%</td>
</tr>
<tr>
<td>Total</td>
<td>372</td>
<td>100%</td>
</tr>
</tbody>
</table>

Amharic being a syllabic writing system, in order to analyze the Damerau-Levenshtein edit distance of the misspellings from their corrections, there is a need to transliterate Amharic characters into Latin-based alphabets. The transliteration is done by following the phonetic mappings of the popular keyboard input methods, Google and Keyman. After the transliteration process, the Damerau-Levenshtein edit distance of the misspellings was computed against their corrections as shown in Table 1. About 78% and 16% of the misspellings are one and two edit distance from their corrections, respectively. That means about 94% of the misspellings have two or fewer edit distances from their corrections.

Amharic being a syllabic writing system, in order to analyze the Damerau-Levenshtein edit distance of the misspellings from their corrections, there is a need to transliterate Amharic characters into Latin-based alphabets. The transliteration is done by following the phonetic mappings of the popular keyboard input methods, Google and Keyman. After the transliteration process, the Damerau-Levenshtein edit distance of the misspellings was computed against their corrections as shown in Table 1. About 78% and 16% of the misspellings are one and two edit distance from their corrections, respectively. That means about 94% of the misspellings have two or fewer edit distances from their corrections.

The Pearson Correlation Coefficient computed between the misspellings and their string lengths is 0.061. This indicates that there is no significant linear correlation between the misspellings and their string lengths.

VII. CONCLUSION

We have developed a manually annotated corpus for Amharic misspellings that can be used to evaluate spelling error detection and correction. The misspellings are categorized as non-word and real-word errors. Besides, the availability of contextual information in the corpus makes it useful in dealing with both types of spelling errors. Moreover, the great majorities of the misspellings are two or fewer Damerau-Levenshtein edit distance away from their corrections; and there is no significant linear correlation between the misspellings and their string lengths.

REFERENCES


