Romanian Google N-grams Filtering Tool

1. BASIC INFORMATION

   1. Tool name

   Romanian Google N-grams Filtering Tool

   2. Overview and purpose of the tool

   This tool is designed to normalize the n-grams contained in the Romanian Google Web 1T Corpus.

   3. A short description of the algorithm

   The algorithm is designed to cope with the following error classes found in the Google corpus by manual inspection:

   - Tokenization errors (e.g. “vizitau.Iulian”, “seara.Chiar”)
   - Spellchecking errors:
     - Romanian words are frequently written with no diacritics at all and as a result, with no context available for disambiguation, there are cases when inserting diacritics will change the meaning of the word, e.g. “fata” means ”the girl” whereas “fătă” means ”gave birth”;
     - Romanian words are quite often written with mixed Romanian ortography using either “â” or ”î” or both, not following the current academic norms. For instance,”miine” is actually ”mâine” and ”preântimpinat” is correctly written as ”preîntâmpinat” because “î” is to be used only at the beginning or the end of words or after morphological prefixes such as “pre”. Thus, one can expect virtually any combination of “â” or ”î” with or without their diacritical marks (i.e. “a” or “i”);
     - Foreign words that are incorrectly mistaken for Romanian words written with no diacritics or otherwise wrongly spelled. For instance, the English word “Romania” may be converted to the correct Romanian word “România” while the English word ”member” is incorrectly spellchecked as ”membre” (limbs);
     - Words that are unknown to the spellchecker dictionary might incorrectly be transformed to match a word in the dictionary. Examples include ”Nomia” from “Nokia”, “Bărbie” from ”Barbie” or “lăcățel” from ”Alcatel” (swap ”A” with ”l” and insert diacritics);
     - There is a fair amount of slang words and Romanian diacritic transliteration such as “bajeti” instead of “băieți”, ”tenishi” insted of ”teniși” (”sh” instead of ”ş”) or ”curatzate” instead of ”curățate” (”tz” instead of ”ț”);
   - Foreign words

   The algorithm constructs a “normative corpus” using the unigram file based on the following accept/reject decisions:

   1. **isPunct** verifies if the word is an accepted Romanian punctuation mark or sequence. If the word matches the Perl regex

```
/^(?:[:;.,"'()<>=+_?!%&*”“„@«»©-]|--|\.\.\.|\.\.)$/
```

   it is **accepted**. If not, control is given to the next filter;

   2. **isRoNumber** verifies if the word is an accepted Romanian numeral. If the word matches the Perl regex

```
/^[0-9]+(?:[,.][0-9]+)?$/
```

it is **accepted**. If not, control is passed to the next filter;

3. **isSpecial** verifies if the word is a special Google unigrams token such as the beginning of the sentence (“<S>”), the end of a sentence (“</S>”) and the unknown word (“<UNK>”). If the word is special, it is **accepted** and if not, control is transferred to the next filter;
4. **isMixedCase** checks the word not to contain mixed case characters (e.g. “MyUsEr”). We have empirically found out that these tokens usually denote user names, site names, software/hardware names which, with a few exceptions, are rare. As before, if the word matches any of the following Perl regexes

```
/[a-zăîâșț][A-ZȘȚĂÎÂ]/
/[a-zA-ZăîâșțȘȚĂÎÂ][A-ZȘȚĂÎÂ][a-zăîâșț]/
```

it is **rejected**. Otherwise, control is passed to the next filter;

5. **containsLettersAndNumbers** checks if the word contains both letters and numbers. As with the previous filter, these words usually denote user names (e.g. “User25”), quantities (e.g. 12.433GHz) or simply meaningless tokens (e.g. “MOP-28-5”) which are usually rare. The Perl regex of this filter is

```
/[A-Za-zăîâșțȘȚĂÎÂ]/ && /[0-9]/
```

and if the word matches, it is **rejected**. If it does not match, control is transferred to the next filter;

6. **isForeignWordOrEntity** rejects all words which contain at least a character that is illegal for a Romanian wordform. In this category, we find all email addresses, URLs, foreign words spelled with specific diacritics, words not properly UTF-8 encoded, etc. The Perl regex of this filter is

```
/[^a-zA-ZăîâșțȘȚĂÎÂ.'-]/
```

and every word which matches it, is **rejected**. If not, the next filter is tried;

7. **isNotARoWord** verifies that the word is a legal Romanian wordform, e.g. to begin with a Romanian letter (Perl regex `/[A-Za-zăîâșțȘȚĂÎÂ]/`) or dash “-“ and end with a Romanian letter, dash or period “.” (if it is an abbreviation) and to contain only letters and/or dashes but never two consecutive dashes. If the word does not conform to the above-mentioned rules, it is **rejected**. If it does, it is passed to the next filter;
8. **spellChecker**: if a word made it so far, it is very likely that is a known or unknown Romanian word and it is passed to the Romanian spell checking algorithm (see the next subsection). If this algorithm is able to determine that either this is a legal Romanian wordform (according to its lexicon, the Romanian wordform lexicon) or a misspelled variant of a legal Romanian wordform, the word is **accepted** along with its corrected form if applicable. If there are none or more than one spelling alternatives, alternative filters are triggered (see subsection 2.1 for further details).

2. TECHNICAL INFORMATION

1. Software dependencies and system requirements

   Our tool has the following dependencies:

   - Perl with no external dependencies
   - English and Romanian word form lexicons
   - The Romanian word form lexicon is available through the MetaShare platform[1]

2. Installation

   No installation required

3. Execution instructions

   Use the following execution sequence:

```
$ filtergoogleuni.pl 1-vocab_cs > 1-vocab_cs.filt
$ filtergooglemulti.pl 1-vocab_cs.filt "directory with google ngrams"
$ addgooglemulti.pl newgooglengramsNN.txt "directory with google ngrams"

for each file fileName in "directory with google ngrams" {
      $ mergecounts.pl fileName
}
```

4. Input/Output data formats

   a. Input data formats

   The input format is specific to the Google Web 1T 5-gram corpus

   b. Output data formats

   The output format is specific to the Google Web 1T 5-gram corpus

5. Integration with external tools

   Not aplicable

3. CONTENT INFORMATION

1. a test input file

   Not aplicable

2. the output file

   Not aplicable

3. approximation of the time necessary to process the test input file.

   On a Core I7 980 3.33Ghz CPU, with 16 GB Ram the execution time for the entire Romanian section of the Google Web 1T Corpus took approximately 31 hours.

4. ADMINISTRATIVE INFORMATION

1. Contact

   Radu Ion (radu@racai.ro)

5. RELEVANT REFERENCES AND OTHER INFORMATION

Thorsten Brants and Alex Franz. 2006. Web 1T 5-gram Corpus Version 1.1. Technical report, Google Research

[1] http://ws.racai.ro:9191/repository/browse/romanian-wordform-lexicon/9776c502c9d211e1a3cb00226410db01337727fa83284d7a9c412d6157ddfff5/