A Stemming Algorithm for the Portuguese Language

V. M. Orengo and C. Huyck.
Proceedings of the Eighth International Symposium on String Processing and Information Retrieval, pages 186-193, 2001

Agenda

- Stemming
- Motivation
- Objectives
- The Algorithm
- Evaluation
- Conclusions and Future Work

Stemming

- Definition
  - Process of conflating the variant forms of a word into a common representation (the stem)
- Example
  - presentation, presented, presenting $\Rightarrow$ present
- Assumption
  - Posing a query with “presenting” implies an interest in documents with “presentation” and “presented”

Motivation

- Studies evaluating the validity of stemming for IR reached contrasting conclusions
- Harman 91
  - Examined effects of 3 algorithms on 3 collections
  - Found no improvements on retrieval performance
  - Number of queries with improved performance tended to equal the number with poorer
- Krovetz 93
  - Stemming improved retrieval performance by up to 35% on some collections

Motivation

- Hull 96
  - Some form of stemming is almost always beneficial
  - Overall improvement ranged from 1-3%
  - For many individual queries stemming made a large difference
- These experiments were done in English collections
- Highly inflected languages (such as Portuguese) may benefit more from stemming

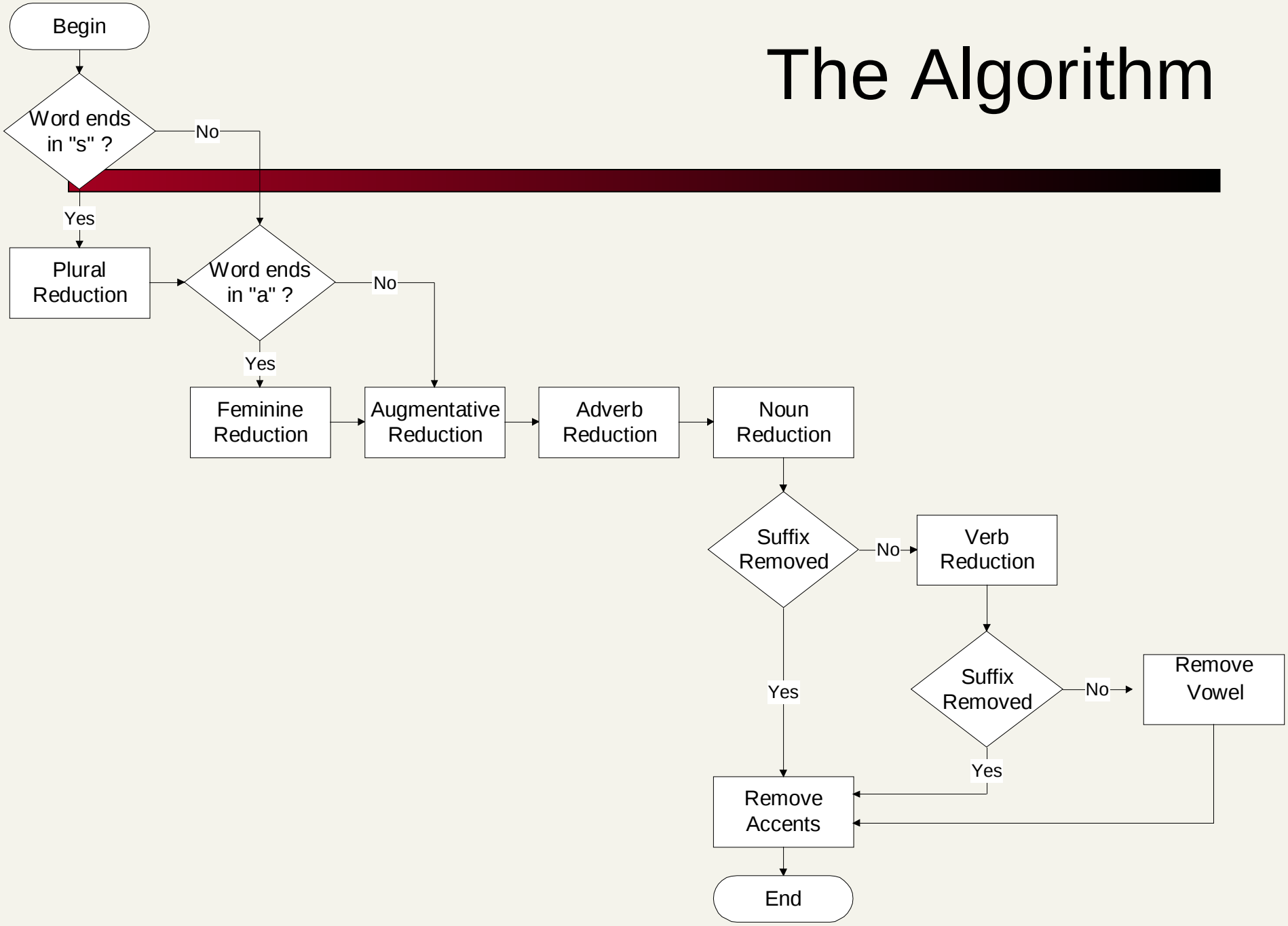
Motivation

- English stemming seems to be a resolved problem
- Porter Stemmer [Porter 80]
  - Simple suffix-stripping algorithm based on rules, without exception lists or dictionary lookups
  - As effective as more elaborated systems
- Similar algorithms have been developed for other languages [Honrado 00, Kraaij 94, Wechsler 97]

Objectives

- Design a suffix-stripping algorithm that is both simple and effective with the target of improving recall, without decreasing precision

The Algorithm



The Algorithm

- Named as “Removedor de Sufixos da Língua Portuguesa” (RSLP)
- Composed by 8 steps
- Each step has a set of rules
- Only 1 rule in a step can apply
- Longest possible suffix is always removed first

The Algorithm

- Each rule states
  - Suffix to be removed
  - Minimum length allowed for the stem
  - Replacement suffix (if necessary)
  - List of exceptions
- Example
  - "inho", 3, ””, {"caminho", "carinho", "cominho", "golfinho", "padrinho", "sobrinho", "vizinho"}

Step 1: Plural Reduction

- Removing the final “s” of the words that are not listed as exceptions
- Not all words ending in “s” denote plural
  - lápis
- Sometimes a few extra modifications are needed
  - bons $\Rightarrow$ bom

Step 2: Feminine Reduction

- Transforming feminine forms to their corresponding masculine
- Only words ending in “-a” are tested
- Not all of them are converted, just the ones ending in the most common suffixes
  - chinesa $\Rightarrow$ chinês

Step 3: Adverb Reduction

- There is just one suffix that denotes adverbs
  - “mente”
- Not all words with “mente” ending are adverbs
- Exception list is needed

Step 4: Augmentative/Diminutive Reduction

- Treat augmentative, diminutive and superlative forms
  - casinha: “inha” is a diminutive suffix
- There are 38 of these suffixes
- Algorithm uses only the most common ones

Step 5: Noun Suffix Reduction

- Tests words against 61 noun (and adjective) endings

Step 6: Verb Suffix Reduction

- Portuguese regular verbs have over 50 forms
- Each one has its specific suffix
- Verbs can vary according to tense, person, number, and mode
- Structure of the verbal forms
  - root + thematic vowel + tense + person
  - and + a + ra + m
- Verbal forms are reduced to their root

Step 7: Vowel Removal

- Removing the last vowel of words not stemmed by steps 5 and 6
  - menino

Step 8: Accents Removal

- Some forms of the word are accented
  - psicólogo e psicologia
- Important that this step is done at this point
- Presence of accents is significant for some rules
  - óis $\Rightarrow$ ol
  - sóis $\Rightarrow$ sol
- If the rule was
  - ois $\Rightarrow$ ol
  - dois $\Rightarrow$ dol (mistake)

Difficulties in Stemming Portuguese

- Dealing with exceptions
  - Not all words ending in “ão” are in augmentative forms
  - RSLP uses exceptions lists
- Homographs
  - casais: “couples” or 2nd person plural of “to marry”
  - RSLP doesn't have information on word categories
  - Different senses of words are not distinguished
    - casais $\Rightarrow$ casal
- Irregular verbs
  - Current version don't treat irregular verbs
  - Less than 1% of the mistakes occur because of this

Difficulties in Stemming Portuguese

- Changes to the morphological root
  - Cases in which the change obeys orthographic rules are being successfully treated
    - ns $\Rightarrow$ m
  - Other cases are not being treated properly
    - emitir $\Rightarrow$ emit
    - emissão $\Rightarrow$ emis
- Proper names
  - As for the Porter stemmer, RSLP stems proper names

Evaluation

- Used a vocabulary of 32,000 words
- Compared RSLP with the Portuguese version of the Porter stemmer
- Used 3 different methods
  - Vocabulary reduction
  - Expected output
  - Paice's method

Evaluation

- Vocabulary reduction
  - Porter: 44%
  - RSLP: 51%
- Expected output
  - Used a corpus with 1,000 manually stemmed words
  - Porter: 71% correctness rate
  - RSLP: 96% correctness rate

Paice's Method [Pace 1994]

- Based on detecting and counting the actual understemming and overstemming errors
- Permits the computation of indexes as
  - Understemming error rate (UI)
  - Overstemming error rate (OI)
  - Stemming weight (OI/UI)
- Involves manually dividing a sample of words into conceptual groups, and referring the actual stemming performance to these groups

Example

- 5 conceptual groups
  1)ajud: ajuda, ajudando, ajudinha, ajudei
  2)duvid: duvido, dúvida, duvidamos, duvidem
  3)chec: checando, chequei, checamos, checou
  4)beb: bebo, bebes, bebi, bebendo, bêbado, bebida
  5)bebê: bebê, bebezinho
- Stemming
  1)ajud, ajud, ajud, ajud
  2)duvid, duvid, duvid, duvid
  3)chec, chequ, chec, chec (understemming)
  4)beb, beb, beb, beb, beb, beb
  5)beb, beb (overstemming)
- UI= 0.088, OI= 0.083, SW= 1.06

Evaluation

- Used 1000 words divided into 170 groups
- Porter
  - UI = 0.215
  - OI = 2.11×10^{-4}
  - SW = 9.81 x 10-4
- RSLP
  - UI = 0.034
  - OI = 9.85×10^{-5}
  - SW = 2.89×10^{-3}

Conclusions

- Development of a Portuguese stemmer
- Simple yet highly effective
- Based on a set of steps composed by a set of rules
- Each rule specifies
  - Suffix to be removed
  - Minimum length allowed for the stem
  - Replacement suffix (if necessary)
  - List of exceptions

Conclusions

- Evaluated using 3 different methods
  - Vocabulary reduction
  - Expected output
  - Paice's method
- Outperformed the Portuguese version of the Porter stemmer in all tests

Future Work

- Using the Portuguese stemmer on an IR system to access its impact over recall and precision

References

- C. D. Paice. An Evaluation Method for Stemming Algorithms. Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, pages 42-50, 1994.