A PUBLIC OPEN-SOURCE ENVIRONMENT FOR A SAFER INTERNET ACCESS

POESIA

TEXT FILTERING FOR SPANISH

Enrique Puertas Sanz

Universidad Europea de Madrid

Contents

- ***Goals***
- Scientific approach
- Design and implementation
- Current results

Goals

- Effective filtering of Spanish text dealing with
  - Pornography
  - Gross language
- Two level filtering (efficiency-driven)
  - Light filtering
  - Heavy filtering

Contents

- Goals
- ***Scientific approach***
- Design and implementation
- Current results

Scientific approach

- Light filter – pornography
  - Statistical text processing
    - Very shallow text analysis
    - Machine Learning
  - High accuracy on "easy" text
  - Efficient

Scientific approach

- Light filter – pornography (details)
  - Very shallow text analysis
    - Basic tokenization
      - Isolating words using separators (space, EOL, etc.)
    - Stop list filtering
      - Filtering out very common words (e.g. Prepositions)
    - Stemming
      - Basic morphology (“analysis”, “analyser” → “analy”)
    - Binary text representation
      - Weight vector (e.g. “sex” occurs → sex has weight 1)

Scientific approach

- Light filter – pornography (details)
  - Machine Learning
    - Filtering tokens with Information Gain
      - Retaining 1% top scoring word stems
    - Support Vector Machines (SVM) & regression
      - SVM linear model

        -1.99 * sex - 0.35 * porn + ... > 0 → safe
      - Logistic regression
        - » Obtain class probabilities by fitting the model

Scientific approach

- Light filter – gross language
  - Swear words in 3 groups (low, med, high)
  - Extracted from the Official Spanish Language dictionary (DRAE), stemmed
  - Operation
    - If any high swear word occurs → score high
    - else if any med swear word occurs → score high ...

Scientific approach

- Heavy filter – pornography
  - More advanced text processing
    - Shallow text analysis with some NLP
    - Machine Learning (as in light filtering)
  - Better accuracy on "difficult" text
  - Less efficient

Scientific approach

- Heavy filter – pornography (details)
  - Shallow text analysis with some NLP
    - Previous approach plus more indicative indexing units
    - Noun Phrases recognition
    - Named Entities recognition ("Pam Anderson" vs. "Bill Gates")

Scientific approach

- Heavy filter – pornography (details)
  - Noun Phrases recognition (3 phases)
    1. Part-Of-Speech tagging training data
       - "el perro come" → "el_det perro_n come_v" where det = determiner, n = noun, v = verb (simplified)
       - Maximum Entropy with MXPOST package 95+% accuracy)
       - Trained on the CRATER corpus (news text)

Scientific approach

- Heavy filter – pornography (details)
  - Noun Phrases recognition (3 phases)
    2. Noun phrases (NPs) as regular expressions
       - E.g. np = det n adj ("el_det niño_n listo_adj")
    3. NP normalization (avoiding tagging incoming text – MXPOST not GPL'ed)
       - Stop list, stemming and ordering
       - E.g. "el niño listo" → "list niñ"

Scientific approach

- Heavy filter – pornography (details)
  - Named Entities recognition
    - As defined in Computational Natural Language Learning (CONLL) 02/03 workshops
      - Named entities = phrases with names of persons, organizations, locations, times and quantities
      - E.g. [PER Wolff] , currently a journalist in [LOC Argentina] , played with [PER Del Bosque] in the final years of the seventies in [ORG Real Madrid] .
    - We partly follow the approach by 02 top performers (Carreras *et al*.)

Scientific approach

- Heavy filter – pornography (details)
  - Named Entities recognition
    - A selection of Carreras text features
      - Focus word capitalization, punctuation marks, etc
    - A number of Machine Learning algorithms
      - Naive Bayes, SVM, kNN, etc.
    - Trained on CONLL Spanish corpora (news text)

Scientific approach

- Heavy filter – gross language
  - Same swear words groups as in light filter
  - Weight vector (3 = high, 2 = med, etc.)
  - Cosine similarity with text input weight vector $\in [0,1] \rightarrow$ score

Contents

- Goals
- Scientific approach
- ***Design and implementation***
- Current results

Design and implementation

- Coded in Java
- Third party (Java) libraries
  - WEKA (learning)
  - HTMLParser (text extraction)
  - Muffin (filtering test)
  - MXPOST (POS-Tagging training data)
- Available at
  - PoesiaSoft/TextFilter/Spanish

Design and implementation

- Package overview
  - indexer (core) – indexing, training
  - gross – gross language
  - ner – Named Entity recognition
  - filter – filtering utils (testing)
  - html2Text – HTML processing and bot
  - main – the filters

Design and implementation

- Statistics
  - Code
    - 50 classes (300 Kb.)
    - 10 data files (10 Mb.)
  - Corpus
    - 35k html files (29k vs. 6k)
    - 1 Gb. of source HTML

Contents

- Goals
- Scientific approach
- Design and implementation
- ***Current results***

Current results

- Official results (beta version, porn light filter)
  - Sample of 4824 Web pages (891/3933)

Predicted Actual	Harmful	Harmless	Total
Harmful	816	75	891
Harmless	4	3929	3933
Total	820	4004	4824
Precision	0.995	0.981	
Recall	0.916	0.999	
F-Measure	0.954	0.990	

Current results

- Official results (beta version, porn light filter)
  - Highlights
    - effectiveness value = 0.916
    - over-blocking value = 0.001

Current results

- Unofficial results
  - Light filter (porn) improved
  - Heavy filter (porn)
    - Slight (untested) improvement due to
      - Bigger feature space
      - NP and NE recognition