TTS Pre-processing Issues for Mixed Language Support

Paulseph-John Farrugia

University of Malta

CSAW '04

Presentation Outline

- Introduction to the relevance of pre-processing for Text-to-Speech (TTS).
- Description of issues specific to the domain under focus, that of SMS messages written in mixed Maltese and English.
- Illustration of techniques with which to address these issues.
- Overview of preliminary implementations and indications for future work.

Introduction

Human Text Processing

Reading Misspelled Text

"*Aoccdrnig to rscheearch at an Elingsh uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, olny taht the frist and lsat ltteres are at the rghit pcleas. The rset can be a toatl mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae we do not raed ervey lteter by ilstef, but the wrod as a wlohe. Fnnuy how the mnid wroks, eh? . . .*"

Human Text Processing

Reading Misspelled Text

"Aoccdrnig to rscheearch at an Elingsh uinervtisy, it deosn't mttaer in waht oredr the ltteers in a wrod are, olny taht the frist and lsat ltteres are at the rghit pcleas. The rset can be a toatl mses and you can sitll raed it wouthit a porbelm. Tihs is bcuseae we do not raed ervey lteter by ilstef, but the wrod as a wlohe. Fnnuy how the mnid wroks, eh? . . . "

- As humans, we can process unstructured and heavily misspelt texts with impressive ease.
- Implementing the same kind of flexibility programmatically is not a simple task.

The Pre-processor

- In a TTS system, the preprocessor provides the first stage of input processing, organising the input text into a standard format that the following modules can process more easily. ([Dutoit, 1997])
- Amongst other things, it is generally responsible for converting numerals and acronyms into their textual interpretation and resolving punctuation.
- Some pre-processing issues are common across input domains (e.g. number, date and acronym handling.)
- Others are specific to the type of input under examination. (e.g. mail message structure handling.)

Mixed Language Support

Pre-processing SMS Messages

Example

"jaqaw bdejt tibza tixel il car? xorta qajjimt il qattusa u issa qed tajjat wara il bieb. bring that btl of last time plz qalbi :)"

- A "rough" domain, which is particularly ill-formatted.
- Generally contains Maltese, English or a mixture of both (65% English, 25% Maltese, 10% other).
- Also contains various shorthands, smileys and spelling errors.
- A real-world system would need to find the means to address these issues in a robust manner.

Code Switching

- As Maltese, we exhibit a tendency to code switch for various reasons.
- A means to classify words as belonging to a particular language is required.
- Use of a lookup dictionary is not sufficient.
- Two main language classification techniques:
  - Use of short word frequencies. [Grefenstette, 1995]
  - Use of n-gram probabilities. [Beesley, 1988, Cavnar and Trenkle, 1994]
- A formalization of the latter approach, appropriate for our purposes, is given.

Bigram Classification

Definitions

- Let **C** be a set of characters.
- Let $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ be a set of n candidate languages.
- For each L_i, let $C_i = \{c_1, c_2, \ldots, c_m\}$, $C_i \subset \mathbf{C}$.
- Let $P_{L_i}(a, b)$ be the probability of bigram ab in L_i text.
- Let $\mathbf{w} = w_1 w_2 \ldots w_k$ be an arbitrary word.

Probability of Word in Language

$$Pw_{L_i}(\mathbf{w}) = \prod_{i=0}^{k} P_{L_i}(w_i, w_{i+1})$$

Bigram Classification

Definitions

- Let **C** be a set of characters.
- Let $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ be a set of n candidate languages.
- For each L_i, let $C_i = \{c_1, c_2, \ldots, c_m\}$, $C_i \subset \mathbf{C}$.
- Let $P_{L_i}(a, b)$ be the probability of bigram ab in L_i text.
- Let $\mathbf{w} = w_1 w_2 \ldots w_k$ be an arbitrary word.

Probability of Word in Language

$$Pw_{L_i}(\mathbf{w}) = \prod_{i=0}^{k} P_{L_i}(w_i, w_{i+1})$$

Bigram Classification

Definitions

- Let **C** be a set of characters.
- Let $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ be a set of n candidate languages.
- For each L_i, let $C_i = \{c_1, c_2, \ldots, c_m\}$, $C_i \subset \mathbf{C}$.
- Let $P_{L_i}(a,b)$ be the probability of bigram *ab* in L_i text.
- Let $\mathbf{w} = w_1 w_2 \ldots w_k$ be an arbitrary word.

Probability of Word in Language

$$Pw_{L_i}(\mathbf{w}) = \prod_{i=0}^{k} P_{L_i}(w_i, w_{i+1})$$

Bigram Classification

Definitions

- Let **C** be a set of characters.
- Let $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ be a set of n candidate languages.
- For each L_i, let $C_i = \{c_1, c_2, \ldots, c_m\}$, $C_i \subset \mathbf{C}$.
- Let $P_{L_i}(a, b)$ be the probability of bigram *ab* in L_i text.
- Let $\mathbf{w} = w_1 w_2 \ldots w_k$ be an arbitrary word.

Probability of Word in Language

$$Pw_{L_i}(\mathbf{w}) = \prod_{i=0}^{k} P_{L_i}(w_i, w_{i+1})$$

Bigram Classification

Definitions

- Let **C** be a set of characters.
- Let $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ be a set of n candidate languages.
- For each L_i, let $C_i = \{c_1, c_2, \ldots, c_m\}$, $C_i \subset \mathbf{C}$.
- Let $P_{L_i}(a, b)$ be the probability of bigram ab in L_i text.
- Let $\mathbf{w} = w_1 w_2 \ldots w_k$ be an arbitrary word.

Probability of Word in Language

$$Pw_{L_i}(\mathbf{w}) = \prod_{i=0}^{k} P_{L_i}(w_i, w_{i+1})$$

Bigram Classification

Definitions

- Let **C** be a set of characters.
- Let $\mathbf{L} = \{L_1, L_2, \ldots, L_n\}$ be a set of n candidate languages.
- For each L_i, let $C_i = \{c_1, c_2, \ldots, c_m\}$, $C_i \subset \mathbf{C}$.
- Let $P_{L_i}(a, b)$ be the probability of bigram ab in L_i text.
- Let $\mathbf{w} = w_1 w_2 \ldots w_k$ be an arbitrary word.

Probability of Word in Language

$$Pw_{L_i}(\mathbf{w}) = \prod_{i=0}^{k} P_{L_i}(w_i, w_{i+1})$$

Character Set Support

- Electronic input devices often do not support: ċ, Ċ, ħ, Ħ, għ, Għ, ż, Ż.
- Unicode should in theory help solve this problem, but in practice it is not in widespread use yet.
- In practice, the following are adopted:
  - A non-standard (font-based) replacement scheme.
  - Adoption of escape sequences representation in ASCII.
  - Replacement with their counterparts: c, C, h, H, gh, Gh, z, Z.
- A “Spell-Checking” Problem
  - Fixed re-write rules: $c \Rightarrow \dot{c}$.
  - Dictionary use: żarbun but not zarbun.
  - Stocasthic/Heuristic re-write rules.

Preliminary Results

Corpora Selection

- In order to estimate the n-gram probabilities, suitable corpora for the languages under consideration is required.
- In order for the probabilities to be meaningful, the corpora are required to be substantially large, and ideally from the same domain as the text that needs to be classified.
- Unfortunately, corpora consisting solely of SMS messages already organised as Maltese or English are not readily available, and deriving substantially sized ones would be a very time consuming exercise.
- An alternative textual corpus is available in the form of the Laws of Malta.

Calculation Frequencies

- The laws are available in PDF format were thus extracted to plain text files.
- The resulting files contained some spurious symbols and characters (such as used for formatting or page numbering). However, given that these made up only a very small percentage of the overall corpora, they would have negligible effect on the overall frequency calculation results.
- For Maltese, two versions were created, one in Unicode and one with the non-ASCII Maltese characters replaced by their ASCII counterparts.
- These were used as the basis for a language classification application that tags whitespace separated tokens according to the maximal $Pw_{L_i}(\mathbf{w})$.

Preliminary Results

- The language classifier was applied to a set of whitespace separated tokens taken from SMS messages with no pre-filtering.
- From hand-checking, this basic, unpolished, process yields a 76% accuracy ratio.
- Analysing the results one finds plenty of room for improvement.
  - No attempt was done to pre-filter the input from non-lexical items, such as smileys and punctuation.
  - Some of the input was in languages other than Maltese and English.

Conlusions & Possible Improvements

- Tendency to fail on short (single letter) words.
- Tagging of 'I' as Maltese – impact of the chosen corpora.
  - Use of dictionary or $|Pw_{Maltese}(\mathbf{w}) - Pw_{English}(\mathbf{w})|$ as a confidence level.
  - Feedback tagged text for self-improvement.
- Tagging of 'u' as Maltese – a more ambiguous situation as it can occur in both languages (in English as short for 'you.') Taking the surrounding context into consideration is necessary in order to resolve the ambiguity.
  - Multi-pass approach using a context window.

For Further Reading

- Dutoit, T. (1997).
  An Introduction to Text-To-Speech Synthesis, volume 3 of *Text, Speech and Language Technology*.
- Grefenstette, G. (1995).
  Comparing two language identification schemes.
- Cavnar, W. B. and Trenkle, J. M. (1994).
  N-gram-based text categorization.
- Beesley, K. (1988).
  Language identifier: A computer program for automatic natural-language identification of on-line text.

Discussion Time