

Predictive Input methods why? How?

Presented by

Anish Patil and Mike Fabian

アニッシュ パティル と マイク ファビアン

Fedora and Red Hat Internationalization Team



Today's Topics

- What are input methods?
- Why predictive input methods are required?
- Theory behind predictive input methods
- Projects that we are working on to implement such input methods



What are input
methods?



Types of input methods

- Character based input methods
Indian, Korean, Vietnamese
- Sentence based input methods
Chinese, Japanese



Types of input methods

- Character based input methods
Indian, Korean, Vietnamese
- Sentence based input methods
Chinese, Japanese



State of input methods





Need

- 1.21 Billion population
 - 74% literate (read & write any language)
 - Still only 5-6% understand English
 - 51% youth in 1.21 Billion
- Diversity in India
 - 22 Officially recognized languages
 - 9 Major scripts



Rest of the world

To preserve endangered languages, the users need good input methods to type them.

- List of extinct language's
 - <http://www.unesco.org/culture/languages-atlas/en/atlasmap.html>



Predictive text

- Statistical techniques
- Probability theory



Language Model

- Lot of words in one language but what is the probability that one word follow other word?
- Simple model: number of occurrence of word/ number of words in the language



Markov Models

- Probability of a word depends only on the probability of a limited history
- Probability of the word depends only on probability of the n previous words
- Unigrams, Biagrams, Trigrams



Example

- Training Set:
 - Start GUADEC is awesome Stop
 - Start GNOME is awesome Stop
 - Start GNOME shell is awesome Stop
- Vocabulary= { Start, GUADEC, is, awesome, Stop, shell }
- Unigram Model:
 - $p(\text{GUADEC}) = 1/16$
 - $P(\text{is}) = 3/16$



- Trigram Model:

- $P(\text{GNOME}/\text{START}, \text{START}) = P(2/3)$

- Whole sentence:

- $P(\text{Start GUADEC is awesome Stop}) =$
 $P(\text{GUADEC}/\text{Start}, \text{Start}) * P(\text{is}/\text{GUADEC}, \text{Start}) * P(\text{awesome}/\text{is}, \text{GUADEC}) * P(\text{Stop}/\text{awesome}, \text{is})$

- $P(\text{Start GUADEC is awesome Stop}) = P(1/3) * P(3/1) * P(2/1) * (3/3)$



Ibus Typing Booster



Available input methods

- Direct Keyboard Input
- Transliteration input methods:
 - Direct keyboard Input
 - Phonetic/itrans
 - Inscript
 - Typewriter/Remington
 - For Latin: Latin-Postfix, Latin-Prefix, Danish-Postfix ...



Ibus typing booster

- Extension to available input methods.
- No need to learn new things.
- The project goal is to improve typing experience and let users enjoy data creation in Indian languages with a boost in typing speed without compromising on data accuracy.



Ibus typing booster

- Supports almost all locales
- Uses hunspell dictionaries for spellchecking
- Supports input methods available in m17n and direct keyboard input



Technology

- Python
- Sqlite



Demo



Drawbacks

- Tied to ibus



libyokan

- Text prediction library written in Vala
- All the key events are handled in library, clients have to just subscribe for text prediction



Need your help

- Testing
- Suggestions for improvements and new features
- Improve hunspell dictionaries
- Creation of free corpora



References

- <http://www.thehindu.com/opinion/editorial/article1599783.ece>
- http://www.wakeupcall.org/our_goal/transforming-india.php
- http://www.tehelka.com/story_main49.asp?filename=Ws050411SECURITY.asp

