

Machine Translation for Indic Languages using Apertium

Pranava Swaroop S
Malaviya National Institute of Technology, Jaipur



Conclusion!

- Most of the Machine Translation Systems are closed
- They rarely are friendly towards their peers
- Most Indian MT systems are stagnant or closed or badly documented
- Interchange is mostly a “Mission Impossible – 5” ;-)
- Integration !!!



Results(Indian perspective)

- Indian MT engines either stagnant
- Undocumented
- Coded by different people
- No uniformity
- Mostly CLOSED
- Mostly Undeployable
- Versioning is a term which most of the Indian organizations(producing MT) have never heard of



So??

- Need for an Open Collaboration
- Would be great if there is a base is readily available
- Must be well documented
- Must be portable
- Must be active



Why all this?

- India has more than 18 Languages defined in the constitution
- Very less literary resources {digital}
- Need for rapid conversion and Immediate generation of digital data
- Need for collaboration
- Though the accuracy may be low during initial phases



And

- Indic Languages belong to the huge group namely:
- Indo-European
- Indo-Iranian
- Indo-Aryan
- Dravidian
- Some of the well known languages from these groups have already well formed corpus and translation rules.



The use?

- Can we inherit some properties?



Any options?

- Apertium!!
- Apertium is an open-source machine translation toolbox (<http://www.apertium.org>) providing:
 - 1 An open-source modular shallow-transfer machine
 - translation engine with:
 - text format management
 - finite-state lexical processing
 - statistical lexical disambiguation
 - shallow transfer based on finite-state pattern matching



- * Spanish–Catalan (apertium-es-ca)
- * Spanish–Portuguese (apertium-es-pt)
- * Spanish–Galician (apertium-es-gl)
- * Occitan–Catalan (apertium-oc-ca)
- * French–Catalan (apertium-fr-ca)
- * English–Catalan (apertium-en-ca)



Do what with that?

- Most of the Indic languages are well known as close neighbours
- Most of the grammatical constructs are almost the same.
- Use apertium for the translation of close neighbours, though it is known that apertium works for sparsely spaced languages



Introduction

- Translation of closely related languages
- The need to write specific translation rules
- <http://apertium.svn.sourceforge.net/viewvc/apert>
- <http://xixona.dlsi.ua.es/~fran/hindidict.txt>



- Please download apertium and It-toolbox from <http://www.apertium.org>
- Live demonstration
- Extend it to different languages
- The application to urdu hindi translation
- <http://sanskrit.uohyd.ernet.in/~anusaaraka/urdu/Urdu-Hindi-Translation/>



Contribute

- Mail me pmadhyastha@acm.org
- Join [apertium@irc.freenode.net](irc://irc.freenode.net/apertium)
- <https://lists.sourceforge.net/lists/listinfo/apertium>

