Developing Punjabi Morphology, Corpus and Lexicon

Muhammad Humayoun[1] Aarne Ranta[2]

[1]Department of Mathematics, University of Savoie, France.
mhuma@etu.univ-savoie.fr

[2]Department of Computer Science and Engineering
Chalmers University of Technology and University of Gothenburg, Sweden.
aarne@chalmers.se

05 - 11 - 2010: PACLIC24

Outline

Outline

Punjabi

- An Indo-Aryan language widely spoken in the Punjab region (in India and Pakistan)
- Written in 2 scripts: *Gurmukhi* and *Shahmukhi*
- Spoken by 88 million people
- As regards the native speakers
  - Most widely spoken in Pakistan
  - 12th or 13th most widely spoken language in the world

Punjabi

- An Indo-Aryan language widely spoken in the Punjab region (in India and Pakistan)
- Written in 2 scripts: *Gurmukhi* and *Shahmukhi*
- Spoken by 88 million people
- As regards the native speakers
  - Most widely spoken in Pakistan
  - 12th or 13th most widely spoken language in the world
- But under resourced
- Resources we report are in *Shahmukhi* script

Motivation

- Punjabi language processing applications and resources only support *Gurmukhi*
- Transliteration is not enough
  - Poor results in existing transliteration systems
  - Different sources for vocabulary
    → *Gurmukhi*: Hindi, Sanskrit, other regional languages
    → *Shahmukhi*: Urdu, Arabic, Persian, other regional languages of Pakistan

Motivation

- Punjabi language processing applications and resources only support *Gurmukhi*
- Transliteration is not enough
  - Poor results in existing transliteration systems
  - Different sources for vocabulary
    → *Gurmukhi*: Hindi, Sanskrit, other regional languages
    → *Shahmukhi*: Urdu, Arabic, Persian, other regional languages of Pakistan

No computational resources for Punjabi written in *Shahmukhi* and spoken in Pakistan

Shahmukhi Script

- Variant of Perso-Arabic script
- Right to left writing
- Optional use of diacritic marks: پَنجَابِی *pañğābī* vs. پنجابی
- Short vowels being not considered as letters of their own but applied above or below a consonant by using diacritics ک vs. (کَ , کِ , کُ)
- Word boundaries not always determined by spaces
  - Space insertion and deletion problem
    شیرِپَنجَاب *šēr-epañğāb* vs. شیرِ پَنجَاب *šēr-e pañğāb*

Contribution

Three important resources released as open-source

`http://code.haskell.org/gf/lib/src/punjabi/`

- **Implementation of inflectional Morphology**
  → In Grammatical Framework (GF)

- **Punjabi Corpus partly from Wikipedia**
  → Containing 0.94 million words (941,284)

- **Semi Automatic Lexicon Extraction**
  → Resulting a lexicon of 13,600 words
  →→ Named entities: 63% (∼8500)
  →→ Surface forms of inflected words: 37% (∼5000)

Outline

Punjabi Morphology

Concatenative inflectional morphology developed in Grammatical Framework (GF)

Grammatical Framework (GF)

- Mainly developed by Aarne Ranta at Chalmers University
- Programming language for NL grammars
- Related to LKB, XLE[1] in purpose
  $\rightarrow$ But based on functional programming and type theory
- Grammar = The Abstract syntax + Concrete syntax

[1] Lexical Knowledge Builder and Xerox linguistics environment

Grammatical Framework (GF)

- Mainly developed by Aarne Ranta at Chalmers University
- Programming language for NL grammars
- Related to LKB, XLE[1] in purpose
  → But based on functional programming and type theory
- Grammar = The Abstract syntax + Concrete syntax
- Abstract syntax: defines semantic conditions to form abstract syntax trees
- Concrete syntax: set of linguistic objects (strings, inflection tables, records) associated to abstract syntax trees
  → rendering and parsing

`http://www.grammaticalframework.org/`

[1] Lexical Knowledge Builder and Xerox linguistics environment

Resource Grammar vs. Application Grammar

On usage, a GF grammar can roughly be divided in:

- **Resource Grammar (GF resource library):**
  - Grammars covering general aspects
    → Morphology + Basic Syntax
  - Linguistic oriented
  - All languages related by a common abstract syntax
  - Has support for an increasing number of languages
    → Complete:14, partial: 8

Resource Grammar vs. Application Grammar

On usage, a GF grammar can roughly be divided in:

- **Resource Grammar (GF resource library):**
  - Grammars covering general aspects
    → Morphology + Basic Syntax
  - Linguistic oriented
  - All languages related by a common abstract syntax
  - Has support for an increasing number of languages
    → Complete:14, partial: 8
- **Application Grammar:**
  - Domain specific
  - May use resource grammars as software API
  - Enables to develop application specific grammars with a very limited linguistic knowledge
  - A number of multilingual applications (Attempto, WebAlt, MathNat, etc) with/without resource library

Morphology in GF

Morphology = Inflection engine + lexicon

Inflection engine

- Part of speech as types
- A set of functions that take a lemma of a word and compute a finite table with all possible word forms
- Each such function represents a paradigm of some part of speech

Lexicon

- Connects the lemma to its paradigm
- a list of word-paradigm pairs

Punjabi Nouns

- Inflects in number (Singular, Plural) and case (direct, oblique, vocative, ablative, locative/instrumental)
- Inherent gender (masculine or feminine)

In GF

```
N = {s : Number => Case => Str ; g : Gender} ;

Number = Sg | Pl ;
Case = Dir | Obl | Voc | Abl ;
Gender = Masc | Fem ;
```

First Paradigm for Nouns

- Nouns are divided into 7 groups based on their inflection[2]
- As a running example, the first paradigm for nouns (mkN01)
- Masculine nouns ending in (ا, aa), (ہ, h), (ع, ’aa) e.g. مُنڈا, munDaa (boy), کوڑا k‘oRaa (horse), بُرقع burk’aa (robe)

	Direct	Oblique	Vocative	Ablative
Singular	munDaa	munDe	munDeaa	munDeoN
Plural	munDe	munDeaaN	munDeo	*non-exist*

	Direct	Oblique	Vocative	Ablative
Singular	burk’aa	burk’e	burk’eaa	burk’eoN
Plural	burk’e	burk’eaaN	burk’eo	*non-exist*

[2]Bhatia (1993:164-166), Shackle (2003:600-601)

First Paradigm for Nouns

N = {s : Number ⇒ Case ⇒ Str ; g : Gender} ;

mkN01 : Str → N ;

```
mkN01 xa = let end = last xa ;
           x = if (end=="ع") then xa else (tk 1 xa)
  in mkN xa (x+"ے") (x+"یا") (x+"یوں") (x+"ے") (x+"یں") (x+"یو")
            ("") Masc ;
```

First Paradigm for Nouns

N = {s : Number ⇒ Case ⇒ Str ; g : Gender} ;

mkN01 : Str → N ;

```
mkN01 xa = let end = last xa ;
               x = if (end=="ع") then xa else (tk 1 xa)
  in mkN xa (x+"ے") (x+"یا") (x+"یوں") (x+"ے") (x+"یں") (x+"یو")
            ("") Masc ;
```

mkN : (x1,_,_,_,_,_,_,x8 : Str) → Gender → N =

```
\sg_dir, sg_obl, ..., pl_dir, pl_obl, ..., gender → {
s = table {
      Sg ⇒ table { Dir ⇒ sg_dir ; Obl ⇒ sg_obl ; ... } ;
      Pl  ⇒ table { Dir ⇒ pl_dir ; Obl ⇒ pl_obl ; ... } ;
    g = gender };
```

First Paradigm for Nouns

Two entries in the Lexicon:

Abstract Syntax: n1_boy : N ;

Concrete Syntax: n1_boy : mkN01 “مُنڈا” (munDaa)

Punjabi Verbs

- Most complex part of speech in Punjabi
- Three classes of verbal sentences in Punjabi:
  → Simple, conjunct and compound (Bhatia 1993:85)

Four paradigms of Simple verbs

1. Having a basic stem form, direct and indirect causatives
   - "nachnaa" to dance, "nachaanaa" made to dance, "nachvaanaa" made to dance by someone
2. Having only basic and direct causative forms
3. Having only a basic and indirect causative forms
4. Having only basic stem form

First and fourth are most common

First paradigm of Simple verbs

The three forms inflect independently in:

- Gender, Number
- Person: 1, 2{casual, familiar, respectful}, 3{near, distant}
- Tense (Subjunctive, Perfective, Imperfective)

And produces 193 word forms

```
V1 = {s : VerbForm1 => Str} ;
VerbForm1 =
    Basic Tense Person Number Gender
    | DirCaus Tense Person Number Gender
    | IndirCaus Tense Person Number Gender
    | Inf | Inf_Fem | Inf_Obl | Ablative | Root | ...
```

Other Part of Speech

- Adjectives
- Adverbs
- The closed classes

Outline

Corpus Development

- Punjabi written in Shahmukhi is difficult to gather on Internet
  - Few people write Punjabi (Shahmukhi)
  - Mostly published in graphic format
- A notable exception is Wikipedia

Two sources for Corpus Collection

1. Punjabi Shahmukhi version of Wikipedia
2. Modern literature Punjabi texts published as graphic images on the website of Academy of the Punjab in North America: http://www.apnaorg.com/
3. Shared by M. G. Abbas Malik in Unicode format

Corpus Data

A corpus of 0.94 million words is made publicly available

		1: Plain	2: Unique sentences	3: Unique words
1.	Wikipedia	327,372	307,047 (93.79%)	22,167 (6.77%)
2.	Literature	651,227	625,111 (95.98%)	46,881 (7.19%)
3.	**Both**	**978,599**	**941,248 (96.18%)**	**51,607 (5.27%)**

Processing Data

1. Ripped off Wikipedia tags, `http://wiki.apertium.org` the modern literature required no processing

Corpus Data

A corpus of 0.94 million words is made publicly available

		1: Plain	2: Unique sentences	3: Unique words
1.	Wikipedia	327,372	307,047 (93.79%)	22,167 (6.77%)
2.	Literature	651,227	625,111 (95.98%)	46,881 (7.19%)
3.	**Both**	**978,599**	**941,248 (96.18%)**	**51,607 (5.27%)**

Processing Data

1. Ripped off Wikipedia tags, `http://wiki.apertium.org` the modern literature required no processing
2. Tokenized the corpora on sentences; multiple occurrences of the sentence were deleted
   - To get rid of Wikipedia template

Corpus Data

A corpus of 0.94 million words is made publicly available

		1: Plain	2: Unique sentences	3: Unique words
1.	Wikipedia	327,372	307,047 (93.79%)	22,167 (6.77%)
2.	Literature	651,227	625,111 (95.98%)	46,881 (7.19%)
3.	**Both**	**978,599**	**941,248 (96.18%)**	**51,607 (5.27%)**

Processing Data

1. Ripped off Wikipedia tags, `http://wiki.apertium.org` the modern literature required no processing
2. Tokenized the corpora on sentences; multiple occurrences of the sentence were deleted
   - To get rid of Wikipedia template
3. Tokenized on spaces, deleting multiple occurrences
   - All non-Shahmukhi characters were deleted

Observations

		1: Plain	2: Unique sentences	3: Unique words
1.	Wikipedia	327,372	307,047 (93.79%)	22,167 (6.77%)
2.	Literature	651,227	625,111 (95.98%)	46,881 (7.19%)
3.	**Both**	**978,599**	**941,248 (96.18%)**	**51,607 (5.27%)**

- Highest frequent words are postpositions, auxiliaries, particles and pronouns
- Wikipedia text normally contains a larger number of named entities then ordinary text?

Explains the reason of having unique words considerably less than the total words

Observations

- Exhibits both space insertion and deletion problems
- Shahmukhi is commonly written without or with a variant number of diacritic marks
- However Wikipedia text mostly without diacritics
- More versions per word with different diacritics
  پَنجَابِي vs. پنجابي
- Tokens with different diacritics are not always same words
  تَیر (to swim) vs. تِیر (arrow)

A fundamental limitation to get a fully vocalized corpus

Wide-Coverage Lexicon

- Building wide-coverage lexicon semi-automatically
- *extract* Tool (Forsberg, Hammarstrom, Ranta 2006)
- The lexicon: extracted by applying *paradigm rules* on corpus
  - Rules: **rule n1 = x+“aa”** {x+“aa” & (x+“e” | x+“eaa” | . . .) }
  - Entries for Lexicon: “n1 munDaa”
- Strictness in rules may effect accuracy and coverage of lexicon. We tried to be in the middle

Wide-Coverage Lexicon

- Building wide-coverage lexicon semi-automatically
- *extract* Tool (Forsberg, Hammarstrom, Ranta 2006)
- The lexicon: extracted by applying *paradigm rules* on corpus
  - Rules: **rule n1 = x+"aa"** {x+"aa" & (x+"e" | x+"eaa" | . . .) }
  - Entries for Lexicon: "n1 munDaa"
- Strictness in rules may effect accuracy and coverage of lexicon. We tried to be in the middle

Manually checked the lexicon from word to word and all incorrect entries were removed resulting in a lexicon of 13,600 words (named entities: 63% ∼8500, lemmas: 37% ∼5000)

Question

Why the lexicon (13,600) is only 26.3% of the unique word count of corpus (51,607)?

- Highest frequent words are postpositions, auxiliaries, particles and pronouns
- Exhibits both space insertion and deletion problems
- Our method is rather simple
  - Tokenization on spaces
  - No context aware queries

Conclusion

Merits:

- Punjabi morphology with fairly good coverage
- GF Supports better abstraction then finite state tools
  Type system and finite functions satisfy completeness
- In future, these components could be utilized for a Punjabi grammar under GF resource library
- But may also adapted to other grammar frameworks

Conclusion

Merits:

- Punjabi morphology with fairly good coverage
- GF Supports better abstraction then finite state tools
  Type system and finite functions satisfy completeness
- In future, these components could be utilized for a Punjabi grammar under GF resource library
- But may also adapted to other grammar frameworks

Limitations:

- A partly vocalized Lexicon
- Corpus tokenized on spaces
- Multi-token words are not yet treated
  - Word boundary recognizer
  - Treated at syntax level

Questions

Thanks!

Are there any questions?

`http://code.haskell.org/gf/lib/src/punjabi/`