Tagging in Indian Languages

Prof. Kavi Narayana Murthy &
Srinivasu Badugu

University of Hyderabad

POS Annotation for Indian Languages: Issues & Perspectives
LDC-IL, CIIL, Mysore
Date:12 & 13-12-2011

Tagging in Indian Languages

- Language, Grammar and Computation
- What is a Word?
- What is a Tag?
- How to design a Tag Set?
- How to Tag a text?
- Examples from Kannada, Telugu

Example

- aMdukee vaaLLu caalaa rakaala samaadhaanaalu ceppaaru.
- aMdukee<ADV-CONJ> vaaLLu<PRO-PER-P3.FM.PL-DIST-NOM> caalaa<ADV-INTF> rakaala<N-COM-COU-N.PL-GEN> samaadhaanaalu<N-COM-COU-N.PL-NOM> ceppaaru<V-TR12-ABS.PAST-P2P3.FM.PL>

- saayaMtraM gaali callagaa viistuMdi.
- saayaMtraM<N-LOC-TIM-NOM> gaali<N-COM-COU-N.SL-NOM> callagaa<ADV-MAN> viistuMdi<V-IN-ABS.PRES.FUT.HAB-P3.FN.SL>

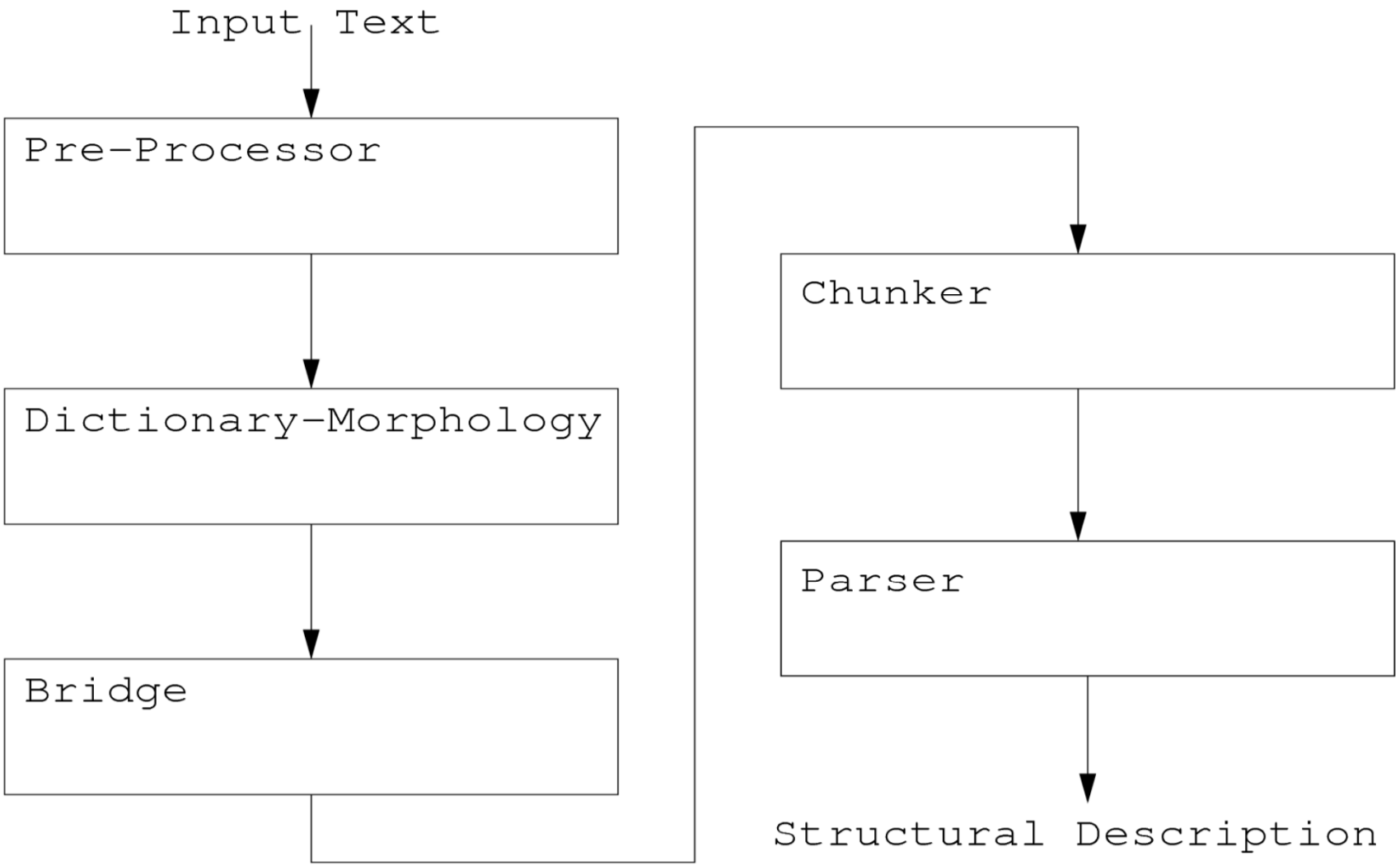
Example

- aa shrama kaalamee daani viluvagaa vuMTuMdi.
- aa<ADJ-DEM> shrama<N-COM-UNC-N.SL-NOM> kaalamee<N-COM-COU-N.SL-NOM-CLIT.ee> daani<PRO-PER-P3.FN.SL-DIST-GEN> viluvagaa<N-COM-COU-N.SL-NOM-ADV.gaa> vuMTuMdi<V-IN-ABS.PRES.FUT.HAB-P3.FN.SL>
- daaraMtoo baTTa tayaaravutuMdi.
- daaraMtoo<N-COM-COU-N.SL-SOC> baTTa<N-COM-COU-N.SL-NOM> tayaaravutuMdi<V-TR12-ABS.PRES.FUT.HAB-P3.FN.SL>

Example

- muDipadaarthaaluu, shrama saadhanaaluu vunnaMta maatraana kuuDaa vastuvu tayaaru kaadu.
- muDipadaarthaaluu<N-COM-COU-N.PL-NOM-CLIT.uu> shrama<N-COM-UNC-N.SL-NOM> saadhanaaluu<N-COM-COU-N.PL-NOM-CLIT.uu> vunnaMta<V-IN-post.RP.adj-CLIT.aMta> maatraana<ADV-POSN||PP-OTH> kuuDaa<ADV-POSN> vastuvu<N-COM-COU-N.SL-NOM> tayaaru<N-COM-UNC-N.SL-NOM> kaadu<V-DEFE-NEG-P3.FN.SL>
- naalugu gaMTalaloo maraNaM saMbhaviMcavaccu.
- naalugu<N-CARD-NHU-NOM> gaMTalaloo<N-LOC-TIM-LOC> maraNaM<N-COM-UNC-N.SL-NOM> saMbhaviMcavaccu<V-IN-INF-aux.permissive>

Architecture



Layered Approach

- Core Grammar plus wrappers
  - Spelling variations, dialects
  - Named Entities, Loan Words
  - Spelling Errors

Corpora

- Kannada
  - TDIL Corpus – About 3 Million Words
  - Hampi Corpus – About 8 Million Words
- Telugu
  - TDIL Corpus – About 3 Million Words
  - LERC-UoH Corpus – About 35 Million Words

Dictionary

- Kannada
  - 58,732 entries, only 5% are ambiguous
  - Headword, grammatical tags, comments only
    - No Meanings
  - Uses a detailed Hierarchical tag set
    - Includes all relevant lexical, morphological, syntactic and semantic features
  - 198 unique tags, 103 tag elements, 92 tag atoms
  - maaDu||N-COM-COU-N.SL-NOM||V-TR1
  - tamma||N-COM-COU-M.SL-NOM::TYPE-kinship||PRO-REF-P23.MFN.PL-GEN
  - aaddariMda||ADV-CONJ::SEE-aadudariMda

Finite State Morphology

- Item and Process Model
  - FSM captures affixes, their sequencing constraints
  - saMdhi processes handle morpho-phonemic changes
  - FSM is bidirectionally used for analysis & generation
- Extended Finite State Machine
  - Finite State Transducer – gives output
  - Category Field, Derivation
  - Handles inflection, Derivation, saMdhi
  - FSM is data, not hard coded

Finite State Morphology

- FSM has 29 states, 401 arcs, 252 tag elements, 172 tag atoms
- It can generate and analyse 18,021 unique word forms with 16,921 unique tags for a given nominal base (2 numbers x 15 case and case-like endings x 220 clitic combinations plus more than 10,000 forms obtained by adding pronominal endings – ex. maneyavaLu)
- It can generate and analyse a mind boggling number of verb forms – difficult even to list

Role of Morphology

- Morphology applies mainly to nouns and verbs
- N and V morphology are disjoint
- Hence n-v ambiguities get resolved by morph
- Morph introduces ambiguities too:
  - maaDi = imperative/conjunctive-participle
  - maaDide : naanu/niinu/adu
  - baMdanoo: dubitative/interrogative
  - maaDee: vocative/emphatic (maaDiyee)
- Yet total ambiguities after morph < 10%
  - Mostly rule governed, syntax can resolve

Morpho-Syntactic Bridge

- Need
- Functions
- Examples

Morpho-Syntactic Bridge

- Combines information pieces from lex and morph and produces tags
- Resolves certain kinds of tag ambiguities
- Handles idiosyncrasies in lex and morph
- Ensures proper mapping to meaning

Tag Set

- Coarse – Fine
- Flat – Hierarchical
- Manual tagging, ML favour flat, course tag sets
- Here we have no restrictions
- More than POS
- Same format used at all levels

Tag Set

- Tags, Tag Elements, Tag Atoms
- First unit is the major category, next one or two are sub-categories, rest are features
- Features appear in more or less the order in which corresponding affixes appear in morph

Main Tags

- V
  - IN/TR/BI/DEFE
- N
  - COM/CARD/LOC/PRP
    - PER/LOC/ORG/OTH
- PRO
  - PER/INTG/REF/INDF
- ADJ
  - DEM/QNTF/ORD/ABS
- ADV
  - MAN/CONJ/PLA/TIM/NEG/QW/INTF/POSN/ABS
- INTJ, SYMB, CONJ
    - SUB/COOR

Tag Definitions

- Each major category/sub-category is precisely defined in terms of lexical, morphological, syntactic and semantic properties
- Ex. Both nouns and pronouns take number and case inflections, both can act as subj/obj and so on. Why should we make a distinction?

Tag Definitions

- That pronouns stand in place of a noun is too vague a definition. Pronouns differ from nouns in significant ways in syntax:
  - The boy runs. He runs. *The He runs.
  - Three boys run. *Three they run
  - That boy runs. *That he runs.
  - The tall boy runs. *The tall he runs.
  - Teenage boy. *Teenage he
  - Bus pass. *It pass
  - Also, Pronoun morphology is usually quite idiosyncratic.
  - Also, pronouns have a distinct role in discourse. Ex. References
  - Therefore, we need to distinguish between n and pr

Tag Definitions

- Similarly, we need to distinguish between common nouns and proper nouns since they behave differently in morphology and syntax:
  - Proper nouns rarely occur in plural
  - Proper nouns are usually not modified, nor do they modify.

Tag Definitions

- Do we need to distinguish between place names and person names?
  - Place names: rarely subj/obj, rarely in accusative, nominative, dative., locative are more common, can modify common names (Hyderabad university, Delhi police) and person names (Guntur Bharadwaj), ...
- Tag set should be designed for a particular purpose. Ex. For Syntactic Parsing. Then lexical, morph, chunking/parsing differences should define the tag set.

Tag Examples

- manege||mane||N-COM-COU-N.SL-DAT
- maaDuttaane||maaDu||V-TR1-PRES-P3.M.SL
- maaDidare||maaDu||V-TR1-PAST-COND
- maaDabeekaagibaMdaaga||maaDu||
  - V-TR1-INF-CMPL-AUX.aagu-CJP.PAST
  - -AUX.baru-PAST-RP-adj-CLIT.aaga
- maaDibiTTaraMtaa||maaDu||V-TR1
  - -CJP.PAST-AUX.biDu-PAST-P3.MF.PL
  - -CLIT.aMte-CLIT.INTG

Tagging Performance

- Assigns tags to more than 90% of words in any given file
- Only about 10% may be ambiguous
- Correctness can be guaranteed
- Ambiguities can be resolved
  - Manually
  - Through Machine Learning
  - Through Syntax
  - Through heuristics

Syntactic Parsing

- Chunking
- Dependency Grammar
- Flexible, Permissive
- Arguments handled top-down, non-arguments are handled bottom-up
- Only sample grammar implemented as yet

Telugu

Corpora

- LERC-UoH Corpus: About 35 M Words
- TDIL Corpus: About 3 M Words - 6,45,410 word types
- Most frequent words in TDIL corpus:

ii	37329
aa	25184
oka	20253
kuuDaa	13101
ani	12847
reMDu	8068
kaani	7477
idi	7096
adi	6935
tana	6878

Corpus Self Coverage

Coverage(%)	Number of Word types
10	61
20	283
30	913
40	2408
50	5704
60	**13427**
70	33043
80	89980
90	289852

Dictionary

- 45,300 entries – less than 3% are ambiguous
  - 274 unique tags, 143 tag elements, 121 tag atoms
  - Ambiguity:
    - 4 tags: Only One Word (0.002%)
      - tarugu||ADJ-ABS||N-COM-UNC-N.SL-NOM||V-IN||V-TR12
    - 3 tags: 28 words (0.06%)
      - laagu||N-COM-COU-N.SL-NOM||V-IN||V-TR12
    - 2 tags: 991 words (2.2%)
      - vaaDu||PRO-PER-P3.M.SL-DIST-NOM||V-TR12
      - paMDu||N-COM-COU-N.SL-NOM||V-IN
    - **1 tag: 44280 words (97.7%)**

Finite State Morphology

- FSM has 19 states, 357 arcs, 253 tag elements, 259 tag atoms
- It can generate and analyse a mind boggling number of verb forms – difficult even to list

Experiments

- First we select 15000 most frequent word forms in TDIL corpus, which give a coverage of over 60% of the corpus
- This list includes the most complex and confusing cases
- These are tagged and manually checked:
  - Total words forms analysed: 15000 (100%)
    - Found in Dictionary: 6483 (43%)
    - Analysed by Morph: 8628 (57%)
    - Total Time Taken 19 min
    - Time per Word 0.0756 Seconds

Analysis

- Root level analysis:
  - Total number of roots: 6693 of which only 318 are ambiguous
  - Category Break-up:
    - Nouns: 4362
    - Verbs: 569
    - Pronouns: 152
    - Adjectives: 981
    - Adverbs: 533
    - Conjunction: 16
    - Interjection: 23

Analysis

- Morph level analysis
  - Ambiguity: 665 words have more than one morph analysis
  - Category Break-up:
    - Nouns: 9227
    - Verbs: 3409
    - Pronouns: 515
    - Adjectives : 1146
    - Adverbs: 600
    - Conjunction: 34
    - Interjections: 23

Types of Ambiguity

- Causative form of verbs: iMcu
  - Ex: kanu to bear, give birth to, to see, observe,
  - Ex: kanipiMcu to seem, appear, be visible
- Verb Reflexive form: konu
  - Ex: paDu to fall,
  - Ex: paDukonu to lie down
- Clitic ee and relative participle ee
  - Ex: uMDee -> uMDa+ee or uMDu+ee
- Accusative case and PNG markar (nu)
  - Ex: perugunu(curd or grow)
- Defective verbs
  - Ex: leedu existential negative or negative +PNG
- Genitive marker and CJP from (i)

Tagging

- Next we have performed POS tagging on different corpora
- F1 is a randomly selected file from TDIL corpus
  - 365 sentences, 4910 words, 2109 word types
  - Sentence length: Min: 2 Max: 45 Avg : 13.5 words
- F2 is set of sentences extracted from the TDIL corpus containing only some 15,000 most frequent words from the corpus.
  - 15100 sentences, 76348 words
  - sentence length: Min: 1 Max: 26 Avg : 5.05 words

Tagging

- F3 from the Eenadu Telugu Newspaper Corpus
  - 33 sentences, 282 words
  - Sentence length: Min: 2 Max: 20 Avg : 8.81 words
- F4 from the Eenadu Telugu Newspaper Corpus
  - 27 sentences, 237 words
  - Sentence length: Min: 2 Max: 28 Avg : 7.8 words

Tagging Performance

File	#sent	#wds	Dict	Morph	UNK	M-AMB	D-AMB	TOT-AMB	Time
F1	365	4910	2186	2389	313	158	170	328	0:6m
			(45%)	(49%)	(6%)	(3%)	(4%)	(7%)	0.07(w/s)
F2	1500	76004	45225	31103	20	4917	3194	8111	1.13m
			(59%)	(41%)	(0%)	(6%)	(4%)	(10%)	0.06(w/s)
F3	33	282	107	173	2	15	10	25	1m
			(38%)	(61%)	(1%)	(5%)	(4%)	(9%)	0.07(w/s)
F4	27	237	88	99	50	8	7	15	1m
			(37%)	(42%)	(21%)	(3%)	(3%)	(6%)	0.08(w/s)

Caching for Speed

- We use 15000 most frequent words forms in TDIL corpus which cover more than 60% as a software cache:

File	Without Cache	With Cache
F1	6min (0.07sec/wd)	3min 6sec (0.04sec/wd)
F2	1h13m (0.06sec/wd)	1min 9sec (0.0009sec/wd)
F3	1min (0.07sec/wd)	17sec (0.06sec/wd)
F4	1min (0.07sec/wd)	15sec (0.07sec/wd)

- To tag the whole TDIL -Telugu corpus, we will need about (0.04*33,00,000) ~ 37hr on a normal desktop PC

Resolving Ambiguities

- Some ambiguities are resolved by Morph:
  - Noun-verb, adj-noun, pronoun-verb:
    - Morph: vaaDavaccu<vaaDu:PRO-PER-P3.M.SL-DIST-NOM||V-TR12:%v-INF-aux.permissive-%--
      - vaaDavaccu<V-TR12-INF-aux.permissive>
    - aDigaaru<aDugu:ADJ-ABS||N-COM-COU-N.SL-NOM||V-TR12:%v-ABS.PAST-P2P3.FM.PL-%--
      - aDigaaru<V-TR12-ABS.PAST-P2P3.FM.PL>
- Some ambiguities are passed-on by morph
    - Morph: paMpee<paMpu:N-COM-COU-N.SL-NOM||V-TR12:%v-FUT.HUB.RP-%--/paMpu:N-COM-COU-N.SL-NOM||V-TR12:%n-SL-obliq-NOM-CLIT.ee-%-->

Resolving Ambiguities

- Genuine Ambiguities:
  - mokkaku<mokka:N-COM-COU-N.SL-NOM:%n-SL-obliq-DAT-
  - mokku:N-COM-COU-N.SL-NOM||V-TR12:%v-NEG.prohibitive-%-->
- Some ambiguities are resolved at chunking level:
  - diini paMDu rucigaa uMTuMdi.
  - diini<PRO-PER-P3.FN.SL-PROX-GEN> paMDu<N-COM-COU-N.SL-NOM||V-IN> rucigaa<ADV-MAN> uMTuMdi<V-IN-ABS.PRES.FUT.HAB-P3.FN.SL>

Resolving Ambiguities

- Some ambiguities are resolved at parsing level:
  - vaaDu pustakaM konnaaDu.
  - vaaDu<PRO-PER-P3.M.SL-DIST-NOM||V-TR12> pustakaM<N-COM-COU-N.SL-NOM> konnaaDu<V-TR12-ABS.PAST-P3.M.SL>
  - vaaLLu giita daaTi loopalaku veLLaaru.
  - vaaLLu<PRO-PER-P3.FM.PL-DIST-NOM> giita<N-COM-COU-N.SL-NOM||N-PRP-PER-F.SL-NOM> daaTi<V-TR12-CJP> loopalaku<N-LOC-PLA-DAT> veLLaaru<V-IN-ABS.PAST-P2P3.FM.PL>

Conclusions

- No Manual tagging, No training data, No ML
- Detailed, Hierarchical Tagging
- About 40% of word forms are directly found in the dictionary, 50-60% are analyzed by morph. Mostly correct.
  - Only some 5-10% of the words remain untagged, unless the input file contains a heavy dose of proper names, loan words, compounds and external saMdhi
  - Separate efforts are on to handle NEs and Spelling Errors.
- One time effort, useful for many applications, easy to maintain and improve, stand-alone system

Plans

- Check and Validate the Dictionary fully
- Check and perfect the morph, tagging modules
- Implement and test chunker
- Design, implement and test parser
- Graphical User Interfaces
- Documentation
- Release

Thank you

- Books of General Interest by the same author:
  - **Freedom** (forthcoming)
  - **Ahimsa** (ask for a copy)
  - **brahmacarya** (down-loadable from 202.41.85.68)

- Contact: **email:** ***knmuh@yahoo.com***
- ***Web: 202.41.85.68***