Morphological Disambiguation of Classical Sanskrit

Oliver Hellwig, University of Düsseldorf

Structure

- Linguistic background, corpus
- System and algorithm
- Improving the morphological analysis
- Outlook and summary

Historical settings

Vedic Sanskrit (1500?-500? BCE)

Vedas, Brahmanas, early Upanishads

Panini (350 BCE? North-West India)

Classical Sanskrit (after Panini)

Is Sanskrit relevant and interesting?

- Biggest (?) corpus of premodern texts
- Reflects „elitist“ Brahmanical wordview
- Broad range of topics: Religion, philosophy, science (medicine, mathematics, …), poetry, epic and narrative literature

Linguistic peculiarities of Sanskrit

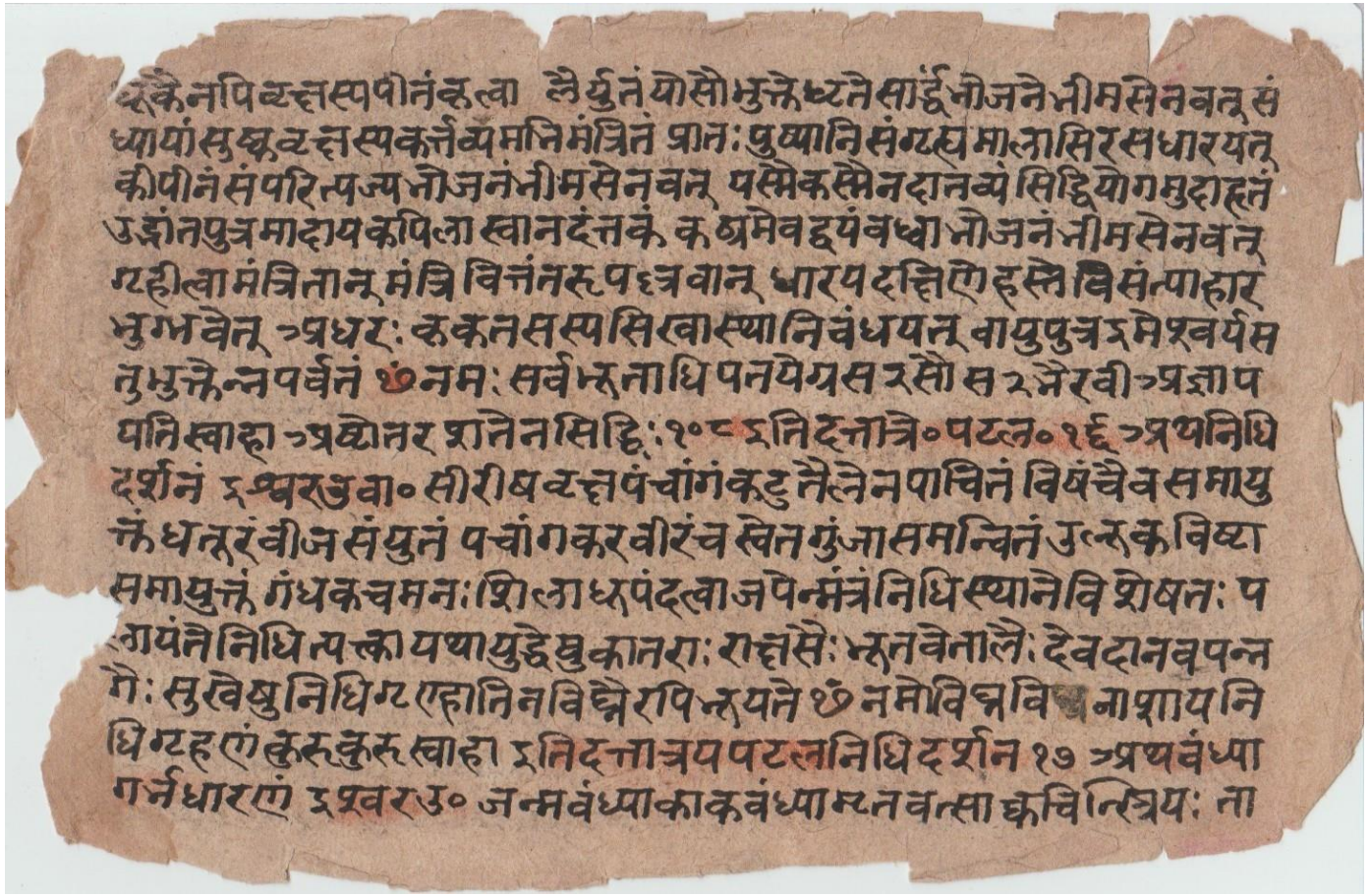
- Noun morphology:
  - 3 genders, 3 numbers, 8 cases: *aśva* ("horse", a masc.): *aśv-aḥ* (nom. sg.), *aśv-am* (acc. sg.), ... *aśv-ābhyām* (ins./abl. du.), ... *aśv-eṣu* (loc. pl.) ...
  - Different inflection classes: aśv-a (a masc.), sīt-ā (ā fem., "name of a woman"), uṣṇih (cons. fem., "a meter"), ...
- Verb morphology:
  - Present stem (ten different classes), future, perfect, aorist; finite verbal forms (incl. absolutive)
  - gam ("to go", 1. present class): gacch-āmi (1. sg. pres.), gam-iṣyāmi (1. sg. fut.), a-gam-am (1. sg. thematic aor.), gatvā (absolutive), gata (past participle), ...

Problems …

- Sandhi: Combination of adjacent phonetic units
  - aśvasy**a**+**a**yanam ("walking of the horse") [rule: **a**+**a**=**ā**] => aśvasy**ā**yanam
  - aśvasy**a**+**ā**hāraḥ ("food of the horse") [rule: **a**+**ā**=**ā**] => aśvasy**ā**hāraḥ (could also be: aśvasy**a**+**a**hāraḥ, "the non-catcher of the horse“, overgeneration of the analyzer!)
- Compounding
  - dvandva (enumeration): hastyaśvoṣṭr-āḥ (<= hasti ("elephant") + aśva + uṣṭr-āḥ ("camels"), "elephant(s), horse(s) and camel(s)")
  - tatpurusha (relation): rājaputr-aḥ (<= rāja ("king") + putra ("son"), "son of the king"); gender = gender of putra (masc.)
  - bahuvrihi (possession): rājaputr-ā strī ("a woman who has a son who is a king"); gender = gender of strī (fem.)

More problems …

- Word order
- Size of the lexicon
- Orthography and ungrammaticality
  - Western style: *yas tv ekāgre cetasi sadbhūtam arthaṃ pradyotayati …*
  - Traditional style: *yastvekāgrecetasisadbhūtam arthaṃpradyotayati*
  - Any intermediate level: *yas tvekāgre cetasi sadbhūtamarthaṃ pradyotayati*



System

- Lexical database with ~ 150.000 lemmata and connections into a semantic inventory
- Corpus: ~ 4.000.000 gold annotated items (lexical and morphological level)
- Linguistic models and information: Sandhi rule base, language models (<- corpus), prebuilt verbal forms
- Tag set
- Linguistic processor

Tokenization

- Split sentence into words
- Try to tokenize words using Sandhi rules:
  - Source string: *āgam*
  - No affix: *āgam* => 1./2./3. sg., root aorist of *ā-gam* ("to arrive")
  - *āga+m*: No solutions
  - *āc*[after Sandhi]+*am* => [compound form of a gramm. term, *āc*] + [a Mantra, *aṃ*]
  - *ā+gam* => "to [ā] the goer [g-am]“
  - *ā+agam* => "to [ā] the tree [ag-am]“
  - *a+agam* => “*the non-tree“
  - *a+āgam* => (bahuvrihi) “*(a person,) who has no singing“
- Viterbi decoding for finding the best path through the graph of hypotheses

Tokenization: Evaluation

	Number of edits			
	0	1	2	≥ 3
≤ 5	14.47	0.19	0.26	0.04
$6-10$	75.63	2.91	1.43	0.28
$11-15$	3.58	0.16	0.17	0.01
≥ 16	0.75	0.04	0.04	0.03
$\sum$	94.43	3.3	1.9	0.36

Morphological analysis: Challenges

- Second step: Choose the most probable morphological analysis for the items in the best lexical path.
- Relevant for approximately 42% of all tokens

aśvasy**a**+**a**yanam ("walking of the horse")

aśvasya (gen. sg.) → ayanam (nom. sg. neutr.)
→ ayanam (acc. sg. neutr.)
→ ayanam (voc. sg. neutr.)

Select the most probable solution!

Morphological analysis: Models

- Original implementation (tri): Viterbi decoding with trigrams of morphological tags. Ignores lexical information!
- Requirements for a better decoding algorithm:
  - Handles categorical data (lexical and morphological information)
  - Sequential?
- Tested:
  - Conditional Random Fields (sequential)
  - Maximum Entropy (non-sequential)

Morphological analysis: Features

- Lexical and morphological information about the target word and all words with a maximal distance of 3 from the target word

aśvasya (gen. sg. masc) → ayanam (nom. sg. neutr.)
→ ayanam (acc. sg. neutr.)
→ ayanam (voc. sg. neutr.)

Features for ayana-:

1. Lexical:
   L_{-2}=…,
   L_{-1}=aśva,
   L_0=ayana,
   L_{+1}=…
2. Morphological:
   M_{-2}=…, M_{-1}=gen.sg.m.,
   M_0=(nom.sg.n.|acc.sg.n.|voc.sg.n.),
   M_{+1}=…

Morphological analysis: Training

- Pre-filtering: Sentences with more than 2 and less than 20 lexical gold items: S_1.
- Use only those sentences from S_1 for which the lexical silver analysis is identical with the lexical gold analysis: S_2
- Training set: 95% of S_2, test set: 5%. No CV.
- Only keep lexical and morphological features that occur with a minimal frequency in the training data.

Morphological analysis: Results (I)

No. of solutions	Proportion	tri	crf	me	fallback	majority
1	58.04	-	-	-	-	-
2	15.74	92.04	93.12	87.56	93.79	93.22
3	9.09	82.48	88.52	82.1	88.56	87.43
4	9.38	77.89	82.94	79.15	82.78	82.56
5	2.98	89.56	91.42	87.18	91.85	91.65
6	1.76	85.82	89.8	82.76	90.76	88.84
7	0.69	85.7	89	86.55	88.88	88.63
8	0.25	76.49	83.44	78.15	84.77	79.14
9	1.51	83.03	84.21	75.03	85.39	83.54
≥ 10	0.54	84.45	86.47	80.72	88.18	85.54

Morphological analysis: Results (II)

		crf			tri			me		
Tag	Prop.	P	R	F	P	R	F	P	R	F
Co.Msc.	14.34	98.52	**99.62**	**99.07**	**98.8**	97.38	98.08	93.54	98.88	96.14
Nom.Sg.Neu.	11.22	**80.92**	**88.07**	**84.34**	73.2	79.09	76.03	75.83	84.29	79.84
Acc.Sg.Neu.	9.11	**81.94**	**76.24**	**78.99**	72.44	69.13	70.75	71.54	74.7	73.09
Nom.Pl.Msc.	7.00	93.91	**98.36**	**96.08**	**94.4**	95.54	94.97	89.79	95.34	92.48
Acc.Sg.Msc.	4.47	**84.4**	79.87	**82.07**	83.08	**80.01**	81.52	75.14	75.96	75.55
3.Sg.Past	3.09	98.9	**99.8**	**99.35**	**99.28**	99.41	99.34	93.51	97.65	95.54
Gen.Sg.Msc.	2.90	90.02	**97.02**	**93.39**	**92.36**	93.96	93.15	89.02	91.67	90.33
Loc.Sg.Neu.	2.86	92.21	89.09	90.62	**93.51**	**90.29**	**91.87**	86.74	85.64	86.19
Nom.Sg.Msc.	2.76	**92.66**	**96.65**	**94.61**	92.31	92.71	92.51	89.41	91.69	90.54
Loc.Sg.Msc.	2.44	85.25	91.09	88.07	**87.57**	**91.25**	**89.37**	82.87	83.83	83.35
3.Sg.Pres.	2.44	**98.28**	**99.09**	**98.68**	96.22	98.68	97.43	92.78	97.53	95.1

Perspectives (I)

- Frame semantic labeling, „Education_teaching“; F scores

	CRF; lex., morph.	CRF; lex., morph., word sem.	Elman; neural embeddings, morph.	Bidir. LSTM; neural embeddings, morph.
Student	3.51	5.26	13.58	**47.24**
Subject	20.44	45.12	43.90	**70.69**
LU	28.78	43.87	78.07	**92.06**
Teacher	8.33	16	15.15	**40.0**

Increasing „neurality“!

Perspectives (II)

- Task: Joint Sandhi resolution and compound splitting using only phonetic information. No external lexical and morphological resources.
- aśvasyāyanam => aśvasya+ayanam. Features: a, ś, v, a, s, y, …
- Bidirectional LSTM with 1-hot-encoding of phonemes as input and softmax output
- Accuracy: 93.2% (vs. 94.4% of the presented system)!