Morphological Disambiguation for Tocharian

Interdisciplinary Project Report

by Gabriel Breiner supervised by Hannes Fellner, Gabor Recski, Judit Acs and Bernhard Koller
Overview

- Introduction
- What is Tocharian?
- The Problem
- Approach
- Evaluation
- Results
- Conclusion
Introduction

- Master’s Student in Data Science
- Interdisciplinary Project
  - “Solve a practical problem in interdisciplinary project work”
  - Problem should not be primarily based in IT or Mathematics.
- Got in contact with Project “Tarim Brahmi”
  - “allow the comprehensive paleographic investigation … ” [of Tocharian]
  - by linking relevant data about manuscripts in one place
The Tocharian Language(s)

- Was discovered around 1900
  - along the routes of the Silk Road
  - in the Tarim Basin
- Actually two very similar languages
  - Tocharian A (TA) - found to the east
  - Tocharian B (TB) - found to the west
- 4th to 10th century CE
- Manuscripts are often translations of Sanskrit
  - Buddhist, scientific, administrative documents.
### Indo-European Languages

<table>
<thead>
<tr>
<th>TA</th>
<th>TB</th>
<th>English</th>
<th>Latin</th>
<th>Sanskrit</th>
<th>Persian</th>
</tr>
</thead>
<tbody>
<tr>
<td>känt</td>
<td>kante</td>
<td>hundred</td>
<td>centum</td>
<td>šatām</td>
<td>sad</td>
</tr>
<tr>
<td>pracar</td>
<td>procer</td>
<td>brother</td>
<td>frāter</td>
<td>bhrātr</td>
<td>barādar</td>
</tr>
</tbody>
</table>
Project “Tarim Brahmi”

- Cooperation between University of Vienna & Austrian Center for Digital Humanities and Cultural Heritage
- “... link the text witnesses to their digital facsimiles on the character level and to publish this material together with a TEI-encoded dictionary in an online database” [7]
- “... all quantifiable features of all characters, ligatures and words will be extracted and compared using software tools” [7]
- Currently working on the dictionary
### Current Progress

#### Transcription

... (kš)-

a1 - su řom klyo irasašíši šák kālýmentwam šáltakár : yårk yńāfímune nam poto irasaşunyey p.kšū kāl(pn)k-

a2 - l: yńnap jymrék yástúcaš kăpnlí jámrák yáltuone 1 irasašíši mák niši palhue irasašíši mák (skarn) (šłą)-

a3 - sfeól : támsof yáltuk irasašíšac kurnslel yárkam : irasašíšac irasašíšac trasašíšac : irasašíšac wrašůj(šůj)

a4 - irasašíši mák praksi náš : támno káu irasaşuney p.kšū pruccmo nři pńákam 2 : irasaşuneyo täm(n)čo (ńo)-

a5 - s (prasišču řešidtîrehs lën se sårvarhsidcide bodhsattu sámudrāṃ kápř hmsaim pradhik yes hmsaim -

a6 - li - sårth sambrudvīpāc pě yámuraš spt kláma klukac wram kálk : spt kláma pokena -

b1 - (tkalik spāt kloms lyomām kălk spāt komsā wálts páltwāyo oplāsyo wram oplās oplā kårmm(ąm) (kălkorā)-

b2 - s păř kųrsárwā āt(p)šľasyo rvpnkām tkanā kălk : tmās rākštőšăsșō dvıpam yęx tmās yaksăšăsșō -

b3 - baladvipam yęx tmās ŕtwar-wăkńa ěršăšľ(yč) ołko rūkońčľaś išanās kćők ŕtwar-wăkńa speś(šin)-

b4 - s-řčk klumšasyo sopis ságareś lęnt lăći waśt pąjástăš ēswe empses (n)kš(č)-

b5 - s āšuk kātkorāš ságareś lătăš činđamānč wńār torń kăpšl pońčăm jambuddvipāc e(kro)-

b6 - rře wawčik šlāk škarm || šămpēnami || măski kătkalām kštānčće irasaši sāmuđorā : troiĥñáuk sams(ə)r (ısra)-

... (šşúneyo)

---

#### Inflectional paradigm

**Present**

<table>
<thead>
<tr>
<th></th>
<th>Singular</th>
<th>Active</th>
<th>Middle</th>
<th>Plural</th>
<th>Active</th>
<th>Middle</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third</td>
<td>sātkār</td>
<td></td>
<td></td>
<td>sātkār</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

**Preterite**

<table>
<thead>
<tr>
<th></th>
<th>Singular</th>
<th>Active</th>
<th>Middle</th>
<th>Plural</th>
<th>Active</th>
<th>Middle</th>
</tr>
</thead>
<tbody>
<tr>
<td>First</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Second</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Third</td>
<td>stāk sātkām</td>
<td></td>
<td></td>
<td>sātkar</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

#### Occurrences: 4

**sātkār**

1. A 1 e1 <ąčš> řom klyo irasašíši šák kālýmentwam sātkāra : yårk yńāfímune nam poto irasaşunyey

2. A 2 b2 kăsu řom klyo amoktsăp ělámy kâlyme sātkāra : yårk yńāmāl máskāra pōtal (kro(ŏ)-bș,pal

3. A 37 b6 <bș >-trăš : kăsu řom klyo sātkāra /// <bș>,-ńră kūsmešāl nńmęńč ///

4. A 79 b5 : kăsmwońey yā(ń)u nu păi akhrťăsăšăi sātkar tri sčăpwyčam : 1 (ń)ẹsltenč ře
The Problem
The Problem

- Create a dictionary of all tokens + morphological & grammatical information
  - Lemma
  - Grammatical Tags in multiple categories (case, pos, gender, …)
- Manually annotating each token/type is very time consuming
- Solution: Train predictors for lemma and grammatical information
  - Iterate over the tokens of all documents
- This project was **not:**
  - beating a benchmark
  - comparing different architecture
  - exploring the best parameter setting
“Worten” => “Wort” + [plural, dativ, ...]
Morphological Disambiguation

- Finding the correct grammatical parse for the morphemes of an inflected word
  - commonly includes lemmatization

- Definition & Approach depends on language
  - English: inflective (barely), but very ambiguous
  - Turkish: agglutinative, vast vocabulary, modular

```
dogs → dog(N+pl)
are → be(aux+pres+3+pl)
are → be(pres+2+sg)
saglamlastirmak → saglam/(adj) las(verb+become) tir(verb+caus) mak(noun+nom)
  = “the thing that causes something to become strong”
```
Tocharian Declension

- 10 Cases
  - 4 Primary Cases: Nominative, Genitive, Accusative, Vocative (only TB)
  - 6 Secondary Cases: Perlative, Comitative, Allative, Ablative, Locative, Instrumental (only TA)
- Secondary cases attach to [stem] + [acc.]
- Seems inflective

<table>
<thead>
<tr>
<th>“King”</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>nom.sg</td>
<td>walo</td>
</tr>
<tr>
<td>acc.sg</td>
<td>lānt</td>
</tr>
<tr>
<td>all.sg</td>
<td>lāntāś</td>
</tr>
</tbody>
</table>
Derivational Morphology

- Productive morphemes can stack together
  - Seems agglutinative

<table>
<thead>
<tr>
<th>snai</th>
<th>preposition</th>
<th>“without”</th>
</tr>
</thead>
<tbody>
<tr>
<td>snaitstse</td>
<td>adjective</td>
<td>“poor”</td>
</tr>
<tr>
<td>snaitstsäññe</td>
<td>noun</td>
<td>“poverty”</td>
</tr>
<tr>
<td>snaitstsäññešeše</td>
<td>adjective</td>
<td>“pitiful”</td>
</tr>
</tbody>
</table>

Ultimately “snaitstsäññešeše” is an adjective.
Mode of Disambiguation

\[ \text{disambiguate}(x) = \{\text{lemma}(x), \ p \in \text{pos}, \ c \in \text{case}, \ \ldots, \ v \in \text{voice}, \ \ldots\} \]

**pos** \{noun, adj, verb, uninfl, unkn\}

**case** \{nom, gen, acc, voc*, per, com, all, abl, loc, ins*\}

**gender** \{m, f, n\}

**number** \{sg, pl, du\}

**person** \{1, 2, 3\}

**tense** \{prs, sbj, impf, opt, pret, imp\}

**voice** \{act, mid\}

*language specific*
Data Exploration
## Labeled Data

<table>
<thead>
<tr>
<th></th>
<th>TB</th>
<th>TA</th>
</tr>
</thead>
<tbody>
<tr>
<td># types</td>
<td>11268</td>
<td>3955</td>
</tr>
<tr>
<td># tokens</td>
<td>29298</td>
<td>13220</td>
</tr>
<tr>
<td># documents</td>
<td>7057</td>
<td>1635</td>
</tr>
<tr>
<td>mean doc length</td>
<td>20.4 tokens</td>
<td>33.24 tokens</td>
</tr>
</tbody>
</table>
41.9% of samples are annotated (TB)  
44.4% for TA

<table>
<thead>
<tr>
<th>labeled data (N=11268)</th>
<th>unlabeled data (N=15560)</th>
</tr>
</thead>
<tbody>
<tr>
<td>0.7</td>
<td>0.2</td>
</tr>
<tr>
<td></td>
<td>.1</td>
</tr>
</tbody>
</table>
How complete are the annotations (rows)?
Ambiguity: How many words have multiple parses?

- Influences choice of model
  - low ambiguity: character-level
  - high ambiguity: word-level & context
- Answer: 30.04% (3386)
- What kind of ambiguities?

\[
\text{ṣamāññeşşē} \to \text{ṣamāne }+ \text{nom }+ \text{m }+ \text{adj}
\]

\[
\text{ṣamāññeşşē} \to \text{ṣamāne }+ \text{acc }+ \text{m }+ \text{adj}
\]

\[
\text{ṣamāññeşşē} \to \text{ṣamāne }+ [\text{nom-acc}] + \text{m }+ \text{adj}
\]
Approach
CoNLL-SIGMORPHON [4]

- Challenge in 2018
- Task 1: Inflection
  - given: lemma + tags
  - generate: inflected form
  - with low ($10^2$), medium ($10^3$) and high ($10^4$) numbers of training samples
- On a large number of different languages
  - inflective and agglutinative Ls represented

- Adequate reference for this project
  - *this is the inverse problem*
  - *many languages in the challenge (inflective and agglutinative)*
  - *according to this we have a medium to high number of training samples*
A string generation task

- **Common approach to SIGMORPHON**
  - and well performing (eg. BME-HAS Acs [5])
- **Include the grammatical tags in the target string**
  - treat tags as characters
- **Train (neural) predictor to generate lemma + tags**
  - character by character
  - solves both lemmatization and grammatical tagging

\[
\text{l}+\text{ā}+\text{n}+\text{t}+\text{ā}+\text{ś} \rightarrow \text{w}+\text{a}+\text{l}+\text{all}+\text{sg}
\]

<table>
<thead>
<tr>
<th>“King”</th>
<th></th>
</tr>
</thead>
<tbody>
<tr>
<td>lemma</td>
<td>wal</td>
</tr>
<tr>
<td>nom.sg</td>
<td>walo</td>
</tr>
<tr>
<td>acc.sg</td>
<td>lānt</td>
</tr>
<tr>
<td>all.sg</td>
<td>lāntāś</td>
</tr>
</tbody>
</table>
**Sequence-to-Sequence Model**

- predicts output sequence given input sequence
  - uses some form of RNN
  - enables different input & output lengths
- in this case: encoder-decoder seq2seq model [2]
- common for translation tasks
  - language generation tasks in general
  - often on word-level
  - in our case: character level
Attention

- Extension of encoder-decoder architecture
  - decoder has access to all encoder outputs
  - weighs encoder outputs (=drawing attention to certain elements)
- Advantages
  - counters long sequence bottleneck
  - provides feedback of the model
  - increase performance
Bahdanau Attention [1]

- “Additive attention”
  - as opposed to multiplicative attention (Luong [3])
- computed in decoder step
- Steps
  - calculate alignment from enc_outs and context
  - softmax alignment (= attention_weights)
  - enc_outs * attention_weights
  - concat attended & embedded
Training Process

<table>
<thead>
<tr>
<th>loss</th>
<th>cross entropy</th>
</tr>
</thead>
<tbody>
<tr>
<td>optimizer</td>
<td>adam</td>
</tr>
<tr>
<td>learning rate</td>
<td>1e-4</td>
</tr>
<tr>
<td>batch_size</td>
<td>1</td>
</tr>
<tr>
<td>embedding_size</td>
<td>50</td>
</tr>
<tr>
<td>context_size</td>
<td>100</td>
</tr>
</tbody>
</table>
Evaluation & Results
Evaluation

● Split the per-sample evaluation in to two parts
  ○ Lemma: Levenshtein Distance
  ○ Gramm.Classes: Precision, Recall, Accuracy

● Levenshtein Distance
  ○ count of operations to transform given string into target string
  ○ operations: add, delete, substitute (characters)

● Precision, Recall, Accuracy
  ○ calculated per gram.category
  ○ tuning for recall would be ideal
1.21 / 1.87

Levenshtein Distance
test / val
Classification Metrics (TB val set)

<table>
<thead>
<tr>
<th>tag</th>
<th>precision</th>
<th>recall</th>
<th>accuracy</th>
</tr>
</thead>
<tbody>
<tr>
<td>case</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>gender</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>number</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>person</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>pos</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>tense</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>voice</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

- average precision: 57.2%
- recall: 60.8%
- accuracy: 82.2%
Conclusion

- **Scores are nowhere near perfect**
  - they do not need to be, since they will be checked
  - team was actually positively surprised with the results

- **Benefits of a black box model**
  - Team “Tarim Brahmi” did not need to know much about ML
  - I did not have to learn Tocharian
  - = less interdisciplinary work?

- **Neural networks are hard to debug**
  - have a proper experimental setup
  - do not get carried away

- **Interdisciplinary Work**
  - Communication is very important (vocabulary!)
  - Ideally: Linguists in control up until preprocessing step
  - How much do they need to know? How much do I need to know?
  - I was possibly very lucky considering the data situation
References


[3] Effective Approaches to Attention-based Neural Machine Translation Luong et al. 2015


[6] CeTOM Website https://www.univie.ac.at/tocharian/?home (25.05.2021)