A Web-framework for ODIN Annotation

Ryan Georgi rgeorgi@uw.edu
Michael Wayne Goodman goodmami@uw.edu
Fei Xia fxia@uw.edu
Department of Linguistics, University of Washington

rgeorgi.co/acl16

Introduction

- NLP tools are largely limited to a few **resource-rich** languages.
- **Interlinear Glossed Text (IGT)** is available for thousands of **resource-poor** languages.
- **IGT's** semi-structured format is well-suited to automatic enrichment methods, such as syntactic projection (see below).
- Extracting **IGT** instances from PDF sources is lossy and noisy.

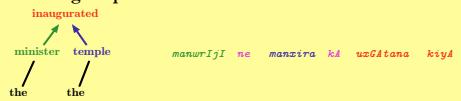
Interlinear Glossed Text (IGT)

LANG	*manwrIjI*	*ne*	*manxira*	*kA*	*uxGAtana*	*kiyA*
GLOSS	minister	[erg]	temple	of	inauguration	did
TRANS	"the	minister	inaugurated	the	temple"	

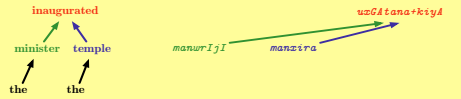
- Unique parallel data source with gloss
- Available for >1,500 languages in ODIN (Lewis & Xia, 2010)

Syntactic Projection

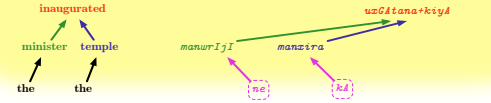
1. Obtain English parse



2. Use word alignment to copy English tree

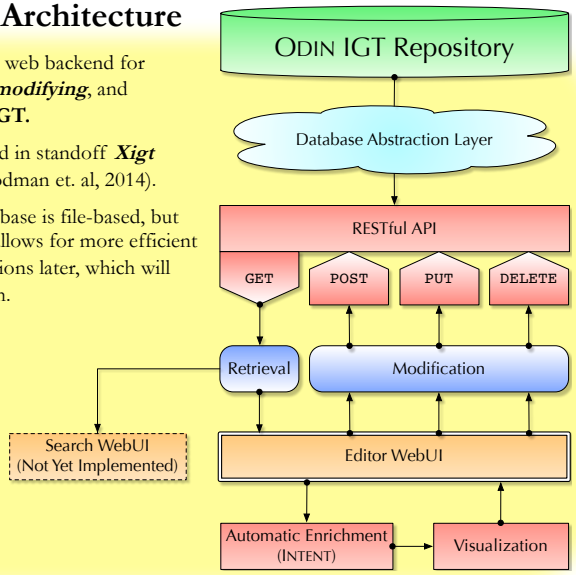


3. Attach Unaligned Words

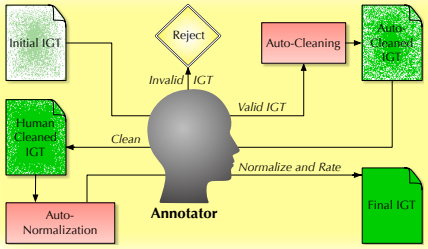


System Architecture

- Developed a web backend for ***accessing***, ***modifying***, and ***enriching*** **IGT.**
- **IGT** is stored in standoff ***Xigt*** format (Goodman et. al, 2014).
- Current database is file-based, but abstraction allows for more efficient implementations later, which will enable search.



Annotation Flow



1. Annotators accept or reject initial **IGT**
2. System attempts auto-cleaning; annotator validates
3. System attempts auto-normalization
4. Annotator validates final IGT instance

Summary

- System allows for rapid cleaning of language data for over 1,500 languages
- Little linguistic knowledge is needed to perform basic cleaning.
- Iterative guidance should result in fewer errors by annotators.
- Visualization of automatic enrichment via **INTENT** (Georgi et. al, 2016) provides immediate feedback on cleanliness of data for annotators.
- Ability of annotators to rate cleanliness of instance will enable improved automatic cleaning approaches.

Future work

- Implement modification of automatic enrichment data.
- POS Tags, Word Alignment, Dependency & Phrase Structures
- Implement Search WebUI for discovering languages and linguistic phenomena.

Acknowledgements

This work is supported by the National Science Foundation Grant BCS-0748919.

We also thank anonymous reviewers for helpful comments.