Evaluating Unsupervised Approaches to Morphological Segmentation for Wolastoqey

Diego Bear and Paul Cook
University of New Brunswick
{diego.bear,paul.cook}@unb.ca

Wolastoqey

- Polysynthetic
- Eastern Algonquian language
- Spoken in regions of what is now:
  - New Brunswick and Quebec, Canada
  - Maine, United States
- Endangered and low-resource language

Morphological Segmentation

- Building block for many language technologies:
  - low-resource language modeling
  - morphologically-aware dictionary search
  - spelling correction
- Finite state models are expensive
- Unsupervised approaches:
  - MorphAGram
  - Morfessor

Datasets and Evaluation

- **Train**
  - 30.1k types from 18.5k example sentences sourced from the Passamaquoddy-Maliseet dictionary
- **Test**
  - **PMLP:** 30 segmented types from Passamaquoddy-Maliseet dictionary
  - **Leavitt-1996:** 102 examples from morphologically annotated text
  - **Leavitt-1996-filtered:** 71 types created by removing particles and preverbs from Leavitt-1996
- We evaluate using boundary P-R-F1 averaged across 10 runs

Experimental Setup

- Train MorphAGram using two grammars
  - Prefixes, Stems and Suffixes
  - Prefixes, Stems, Suffixes and Sub-morphs
- Scholar-seeded (Sch.) and standard (Std.)

Results

PMLP

Grammar	P		R		F1	
Morfessor	0.678		0.377		0.485	
Std. PrStSu	0.619	(0.026)	**0.623**	(0.021)	**0.621**	(0.021)
Std. PrStSu + SM	0.736	(0.021)	0.504	(0.027)	0.598	(0.024)
Sch. PrStSu	0.644	(0.022)	0.571	(0.030)	0.605	(0.025)
Sch. PrStSu + SM	**0.738**	(0.031)	0.466	(0.025)	0.571	(0.026)

Leavitt-1996

Morfessor	**0.710**		0.588		0.643	
Std. PrStSu	0.417	(0.022)	**0.800**	(0.022)	0.548	(0.023)
Std. PrStSu + SM	0.611	(0.021)	0.757	(0.017)	**0.676**	(0.018)
Sch. PrStSu	0.450	(0.025)	0.737	(0.019)	0.559	(0.022)
Sch. PrStSu + SM	0.605	(0.025)	0.747	(0.016)	0.668	(0.017)

Leavitt-1996-filtered

Morfessor	0.668		0.452		0.539	
Std. PrStSu	0.544	(0.025)	**0.668**	(0.021)	0.599	(0.022)
Std. PrStSm + SM	**0.772**	(0.032)	0.616	(0.022)	**0.685**	(0.022)
Sch. PrStSm	0.630	(0.022)	0.617	(0.019)	0.623	(0.018)
Sch. PrStSm + SM	0.763	(0.019)	0.599	(0.020)	0.671	(0.016)

Example Output

Word	Approach	Segmentation
alitahasuwinuwok	Gold standard	ali+tahas+uwin+uwok
	MorphAGram	ali+tahas+uwin+uwok
	Morfessor	al+itahasu+winuwok
kpeciptulonen	Gold standard	k+peci+pt+ul+on+ en
	MorphAGram	k+pecip+t+ul+on+en
	Morfessor	kpeci+ptul+onen
wicihtaqik	Gold standard	wici+ht+aq+ik
	MorphAGram	wi+ci+ht+a+qik
	Morfessor	wici+htaq+ik

Conclusions and Future Work

- **Conclusions**
  - MorphAGram outperforms Morfessor for Wolastoqey
- **Future Work**
  - Evaluate on downstream applications
  - Finite-state morphological analyzer for Wolastoqey