ISP RAS

Development and Integration of Text Processing Tools for Armenian into Texterra

Tsolak Ghukasyan
Head of the Laboratory, RAU

Texterra: Overview

Instruments	Model for Armenian
Sentence, word boundary detection	Completed
Word embeddings	Completed
Part-of-speech tagging	In progress
Spelling correction	In progress
Named entity recognition	In progress
Syntactic parsing	In progress
Lemmatization	In progress
Sentiment analysis	-
Key concepts extraction	-

Main challenge: scarcity of annotated resources.

Texterra: Progress

Sentence, word boundary detection	
Data: Over 20 000 manually processed sentences (fiction, blogs, news)	*Models*: • Regular expressions • OpenNLP's maximum entropy model
Word embeddings	
Part-of-speech tagging	
Spelling correction	
Named entity recognition	

ISP

Texterra: Progress

Sentence, word boundary detection

Word embeddings

Data:
- Wikipedia
- Eastern Armenian National Corpus (EANC)
- News articles, fiction, blogs

Models:
- Word2vec
- GloVe *(analogies: 26.99%)*
- fastText *(analogies: 26.74%)**

*almost x2 better than the model published by Facebook's AI team

Part-of-speech tagging

Spelling correction

Named entity recognition

Texterra: Progress

Sentence, word boundary detection

Word embeddings

Part-of-speech tagging

Data:
- EANC public (manually tagged over 20 000 tokens)
- EANC subcorpus for research
- ArmTreebank

Models:
- Averaged Perceptron (95.3% acc. on EANC)
- Bi-LSTM+CRF

Spelling correction

Named entity recognition

Texterra: Progress

Sentence, word boundary detection

Word embeddings

Part-of-speech tagging

Spelling correction

- Norvig's algorithm
- Norvig's algorithm with phonetic adjustment (acc. gain +7.9%)
- Noisy Channel (acc: 67%) (Kernighan et al. 1990)

Named entity recognition

Texterra: Progress

Sentence, word boundary detection

Word embeddings

Part-of-speech tagging

Spelling correction

Named entity recognition

Data: Dataset extraction from Wikipedia (Nothman et al, 2013)
- *Automated **Wiki** article classification based on **Wikidata***
- *1500 BBN-annotated texts and counting*

Model: Bi-LSTM+CRF

ISP

Texterra: What's next?

Syntactic parsing	
Data: ArmTreebank	*Model:* Google's syntaxnet
Lemmatization	

Texterra: What's next?

Syntactic parsing	
Lemmatization	
Data: - EANC - Wiktionary - ArmTreebank	*Models:* - Rules and lookup table-based - Word vector based

Thank you!