NLP tools for Uzbek language

Informational resources

Thesaurus of the Uzbek language

Text corpus for the Uzbek language

Audio package for Uzbek language

Morphological Analysis

UzMorph – Morphological analyzer for Uzbek language

UzSyntaxParser – Syntactical parser for Uzbek language

Recommended NLP tools, which can be used for teaching

Dependency Parser

MaltParser: A parser based on the shift-reduce method.

MSTParser: A tool for dependency parsing based on maximum spanning trees.

General NLP Libraries

NLTK: A general library for NLP written in Python.

OpenNLP: A library written in Java that implements many different NLP tools.

Stanford CoreNLP: A library including many of the NLP tools developed at Stanford.

Language Modeling

SRILM: An efficient n-gram language modeling toolkit that features a variety of features. A variety of smoothing techniques (including Kneser-Ney), class based models, model merging, etc.

Machine Learning

Classias: A library implementing many different kinds of classifier algorithms, both online and batch.

[Mallet](#): A machine learning package for use in natural language processing. It implements hidden Markov models, maximum entropy Markov models, and conditional random fields. Written in Java.

Machine Translation Alignment

[Berkeley Aligner](#): An alignment toolkit implementing both supervised and unsupervised word alignment models.

[pialign](#): A phrase aligner based on inversion transduction grammars that can create compact but effective translation models.

Machine Translation Decoder

[cdec](#): A parsing-based decoder implementing tree and forest translation.

[Travatar](#): A tree-to-string decoder for syntax-based translation.

Phrase Structure Parsing

[Berkeley Parser](#): A context free grammar parser with models for (at least) English, Arabic, Chinese, Bulgarian, French, and German.

[Stanford Parser](#): A parser that can output both CFG parses and dependencies. Can parse English, Chinese, Arabic, French, and German.

Pronunciation Estimation

[KyTea](#): A toolkit for word segmentation and pronunciation estimation.

[mpaligner](#): A program for aligning graphemes to phonemes for training pronunciation estimation systems, mainly for use with Japanese (site is also in Japanese).

[Phonetisaurus](#): A WFST-based toolkit for grapheme-to-phoneme and phoneme-to-grapheme conversion.

Speech Recognition

[Julius](#): An open-source decoder for large vocabulary automatic speech recognition.

Finite State Models

[Kyfd](#): A decoder for text-processing systems build using weighted finite state transducers.

[OpenFST](#): A library implementing many operations over weighted finite state transducers (WFSTs) to allow for easy building of finite-state models.