Army Research Laboratory

Arabic Natural Language Processing System Code Library

by Stephen C. Tratz

ARL-TN-0609

June 2014

Approved for public release; distribution is unlimited.
NOTICES

Disclaimers

The findings in this report are not to be construed as an official Department of the Army position unless so designated by other authorized documents.

Citation of manufacturer’s or trade names does not constitute an official endorsement or approval of the use thereof.

Destroy this report when it is no longer needed. Do not return it to the originator.
Arabic Natural Language Processing System Code Library

Stephen C. Tratz
Computational and Information Sciences Directorate, ARL
This technical note provides a brief description of a Java library for Arabic natural language processing (NLP) containing code for training and applying the Arabic NLP system described in the paper "A Cross-Task Flexible Transition Model for Arabic Tokenization, Affix Detection, Affix Labeling, POS Tagging, and Dependency Parsing" by Stephen Tratz presented at the Statistical Parsing of Morphologically Rich Languages (SPMR) workshop held in Seattle in conjunction with the Empirical Methods in Natural Language Processing (EMNLP) conference of October 2013. The system is capable of clitic separation, inflectional affix identification and labeling, part-of-speech tagging, and dependency parsing for Arabic. The code, which is extended from previously released graduate student code, also supports English part-of-speech tagging, dependency parsing, and semantic disambiguation tasks. In general, the code library is expected to be of most value to natural language processing researchers.
## Contents

1. Introduction 1

2. File Overview 1

3. Code Compilation 2

4. Training Instructions 2

5. Applying the System to New Examples 2

6. License 3

7. History 3

8. Important Note 4

9. Papers to Cite 4

Distribution List 7
INTENTIONALLY LEFT BLANK.
1. Introduction

This technical note provides a brief description of a Java library for Arabic (and also English) natural language processing (NLP), containing code for training and applying the Arabic NLP system described in Stephen Tratz’s 2013 paper “A Cross-Task Flexible Transition Model for Arabic Tokenization, Affix Detection, Affix Labeling, POS Tagging, and Dependency Parsing” in the 2013 Statistical Parsing of Morphologically Rich Languages (SPMRL) workshop (see section 9 for all relevant publications).

The system is capable of clitic separation, inflectional affix identification and labeling, part-of-speech tagging, and dependency parsing for Arabic. For more information on these topics and the overall system, the reader is referred to the SPMRL 2013 publication.

For historical reasons (section 7), the library also includes components for English part-of-speech tagging, dependency parsing, preposition sense disambiguation, noun compound interpretation, and possessives interpretation. Much of the code is relevant to general purpose NLP applications (e.g., the WordNet interface) and is not specific to either Arabic or English.

2. File Overview

The following is a short description of the top level files and directories.

- build/ - directory with Ant build scripts for running dependency converters, training the various subsystems, and running some experiments
- conf/ - directory with files related to Arabic constituent-to-dependency conversion as well as the English semantic annotation modules
- data/ - directory with a variety of data used by different submodules
- docs/ - directory with additional documentation
- examples/ - a directory with some example input and output files
- jar/ - a directory with a precompiled copy of this code
- lib/ - a directory for any required .jar files
- scripts/ - a directory with some scripts for applying trained models to new data
- src/ - the source code
- README - a README file giving an overview similar to this technical note
LICENSE – the license file
finegrainconverter-0.2.3.jar – English dependency converter jar from http://sourceforge.net/projects/miacp/

3. Code Compilation

The code is typically compiled using Apache Ant, freely available software typically used for building software projects. To download, or for more information, see http://ant.apache.org/.

A variety of ant build XML files are contained in the build subdirectory. To create a jar file containing the compiled code, switch to INSTALLATION_DIR/build and type “ant -f build_common.xml makeJar”.

4. Training Instructions

For the Arabic system, see doc/AR_TRAINING_INSTRUCTIONS.
For the English system, see doc/EN_TRAINING_INSTRUCTIONS.

5. Applying the System to New Examples

Example scripts for applying a trained model to new data are provided under the scripts subdirectory. Short example texts are provided in the examples subdirectory.
6. License

This project constitutes a work of the United States Government and is not subject to domestic copyright protection under 17 USC Â§ 105.

The project includes and derives from code licensed under the terms of the Apache 2.0 License and is, therefore, licensed under the Apache 2.0 License. See the doc/LICENSE-2.0 file for more information. This file may also be found at http://www.apache.org/licenses/LICENSE-2.0.

A copy of WordNet 3.0 (http://wordnet.princeton.edu/wordnet/), a freely available lexical database for English, is included with this project. Its license is provided in the ‘doc/WORDNET_LICENSE’ file.

A modified version of the pennconverter, freely available constituent-to-dependency conversion software described by Richard Johansson and Pierre Nugues in their paper “Extended Constituent-to-dependency Conversion for English,” in the proceedings of the 2007 Nordic Conference on Computational Linguistics, is included with this project. Its license is provided in the ‘doc/PENNCONVERTER_LICENSE’ file.

7. History

This project is derived from a collection of code written as part of my graduate student/thesis work at USC/ISI that was publicly released under the Apache 2.0 license in late 2011 at http://www.isi.edu/publications/licensed-sw/fanseparser/index.html.

In mid-2012, a project derived from the original University of Southern California/Information Sciences Institute version of the software was released at http://sourceforge.net/project/s/miacp/. This newer version used a slightly different English dependency scheme and contained a variety of improvements. However, the PropBank-style SRL module was not maintained.

This current project was derived from the http://sourceforge.net/projects/miacp/ release by NLP researchers at the U.S. Army Research Laboratory. A significant portion of the code has been replaced with newer code, deleted, or otherwise revised. The focus of this effort was to create the system described in the SPMRL 2013 paper mentioned in section 9. However, the English POS
tagger, parser, preposition sense disambiguator, noun compound relation disambiguator, and possessive relation disambiguator are still functional. (The English components may be slightly slower and slightly less accurate than those in the http://sourceforge.net/projects/miacp/ release.)

8. Important Note

This release contains a variety of bug fixes and other generally small changes that have been made since the SPMRL 2013 paper publication and, as such, results will not be identical—in general, the results will be insignificantly better, but there may be some cases where performance is worse.

9. Papers to Cite

Authors wishing to cite to this work should refer to the following papers.

Arabic

English
(Note: These are for earlier versions of the English system/components and, thus, it may make more sense to cite one or more of these along with the 2013 SPMRL paper.)


Dependency Parsing
Preposition Sense Disambiguation


Noun Compound Interpretation

Possessives Interpretation
INTENTIONALLY LEFT BLANK.
<table>
<thead>
<tr>
<th>NO. OF COPIES</th>
<th>ORGANIZATION</th>
</tr>
</thead>
<tbody>
<tr>
<td>1</td>
<td>DEFENSE TECHNICAL INFORMATION CTR DTIC OCA</td>
</tr>
<tr>
<td>2</td>
<td>DIRECTOR US ARMY RESEARCH LAB RDRL CIO LL IMAL HRA MAIL &amp; RECORDS MGMT</td>
</tr>
<tr>
<td>1</td>
<td>GOVT PRINTG OFC A MALHOTRA</td>
</tr>
<tr>
<td>1</td>
<td>DIRECTOR US ARMY RESEARCH LAB RDRL CII T S TRATZ</td>
</tr>
</tbody>
</table>
INTENTIONALLY LEFT BLANK.