11-712: NLP Lab Report

Weston Feely

April 26, 2013

Abstract

This project concerns the creation of an open-source lemmatizer for the Farsi (Persian) language. This lemmatizer is being developed in a finite-state paradigm using the FOMA open source morphological analysis toolkit. The goal of the project is to return the lemma (i.e. dictionary word form) for the surface form of Farsi words. While there have been previous efforts to make lemmatizer or stemmer for the Farsi language, this project aims to create a lemmatizer that can be easily integrated into a larger natural language processing system. The resulting lemmatizer will be included in a Farsi NLP toolkit, which will include a tokenizer, a part-of-speech tagger, this lemmatizer, and a dependency parser.

1 Basic Information about Farsi

Farsi, also known as Persian, is an Indo-European language of the Indo-Iranian branch with over 100 million speakers. It is the official language of Iran, where it is called Farsi, as well as Afghanistan, where it is called Dari, and Tajikistan, where it is called Tajik. There are some differences between these three dialects, so this project will be primarily concerned with the language as written in Iran.

Farsi is an agglutinative language, meaning that most affixes correspond to one syntactic category. Farsi has a productive derivational morphology, creating new words using prefixes and suffixes which attach to word roots, as well as creating new words using compounding. Farsi nouns are marked for number, either singular or plural, and definiteness. Farsi does not have grammatical gender or case marking, except for one optional object case marker. Farsi verbs conjugate into past, present, and future tenses, with additional conjugations for progressive and perfective aspect, as well as subjunctive mood. Farsi is a pro-drop language, meaning pronouns are frequently ommitted from sentences, and the basic word order is Subject-Object-Verb (SOV).

The writing system for Farsi is the Persian alphabet, which is written from right-to-left. The Persian alphabet is based on the Arabic alphabet, with the addition of four letters to represent the sounds /p/,/tʃ/,/ʒ/,/g/, which do not exist in Arabic. The Persian alphabet is an abjad, meaning only consonants and long vowels are written. Letters join to one another, changing shape, with most letters having a beginning, middle, and final form, depending on where in the word the letter is written. Some morphemes do not join with the following letter in their word, in which case a space or a special zero-width-non-joiner character is placed in between the two morphemes.

2 Past Work on the Morphology of Farsi

Recent previous work on Farsi morphology includes the creation of two Farsi morphological resources. The first is a morphological lexicon for Persian, called PerLex (Sagot and Walther, 2010),

which is freely-available online. PerLex consists of a mapping from Farsi lemmas to surface word forms for several thousand words. The second is a stemmer for Persian, called Perstem (Jadidinejad et al., 2010), which is freely available online. Perstem is implemented in Perl and uses a romanization scheme and many complex regular expressions to create stems for Farsi words, suitable for information retrieval and some natural language processing applications. Past work on Farsi morphology also includes a unification-based morphological analyzer by Megerdoomian (Megerdoomian, 2000) which was later adapted to a finite-state morphology paradigm (Megerdoomian, 2004).

3 Available Resources

The two freely-available Farsi morphology resources mentioned above, PerLex (Sagot and Walther, 2010) and Perstem (Jadidinejad et al., 2010), will be utilized in this project. PerLex will provide valuable examples of Farsi lemma and surface word form pairs, which will be especially useful for creating patterns of Farsi morphological derivation for complex part-of-speech categories like verbs. Perstem will also be useful as a tool for comparison with the lemmatizer developed for this project, although the types of lemmas for this project will differ from the stems generated by Perstem, due to differing tokenization and lemmatization schemes.

The corpus for this project will be the Dadegan Persian Dependency Treebank (Rasooli et al., 2011), which contains 30,000 sentences of Farsi text. This corpus is in the format of the CONLL-X shared task for multilingual dependency parsing. The treebank contains gold-standard lemmas for each word in the treebank, which will be useful to check the validity of the lemma hypotheses from the lemmmatizer. The data will be split into three sections: two smaller data sets for development and testing, and one larger data set for initial training. This dependency treebank is being used for this project because a secondary goal of this project is to integrate this lemmatizer into a larger natural language processing system, which includes a tokenizer, a part-of-speech tagger, and a dependency parser trained on this treebank. For this reason, the lemmatization scheme for this project will follow the gold standard lemma style of the Dadegan treebank. Additionally, the Dadegan treebank comes with a verb valency lexicon, which will be a valuable source for the past and present roots of many of the verbs in the treebank.

4 Survey of Phenomena in Farsi

Farsi text poses a challenge for natural language processing tasks, since the Persian alphabet has multiple forms for each letter and letters change forms when joining with their surrounding letters. Also, certain letter pairs are prevented from joining, for stylistic reasons, by the insertion of a space or zero-width-non-joiner character in between the pair.

Derivational morphology in Farsi can be quite complex; a single verbal root can take many affixes which allow the word to become different nouns or adjectives. For example, the root of the verb "to know" can take suffixes that form the nouns "scientist, university, knowledge, wisdom" and the adjectives "scholarly, wise, ignorant". This is due to the agglutinative nature of the language. Recovering the correct lemma for a Farsi word will require the recognition of different affixes for different part-of-speech categories, so part-of-speech tags will need to be included for the lemmatization process.

Farsi verbal grammar also poses a challenge for morphological analysis, since many verb conjugations include multiple space-separated parts in a single token in the Dadegan treebank. These

consist of usually the present or past root of the verb, plus one or more affixes and auxiliary verbs. Lemmatizing Farsi verbs will require recognizing both which part of a verbal conjugation is the main verb, and then the past and present roots of the main verb will need to be returned as the lemmas.

Additionally, Farsi also makes use of many light verb constructions. These are comparable to English light verb constructions such as "take a bath", meaning "bathe", but many more Farsi verbs are of this type. For example, the verb for "to think" is composed of a noun + light verb pair which literally mean "thoughts" and "do". To create a lemma from such a light verb construction, the second part of the construction needs to be recognized as the verb and the lemmas for this verb need to be recovered from the surface form of the verb.

5 Initial Design

The initial design for the Farsi lemmatizer pipeline is as follows:

Input Text → Romanizer → FOMA FSTs → Romanizer → Word Lemma

Here are the details for each module:

1. Input Text: The input text for this project will be Farsi words written in Farsi script and tokenized following the style of the Dadegan treebank tokens. A Farsi tokenizer is being developed concurrently with this project, which will tokenize free text in this tokenization style, as well as providing text normalization (replacement of word-internal whitespace with zero-width-non-joiners, etc.) for open text.

2. Romanizer: This module will romanize the Farsi text into the Latin alphabet, and vice versa. The input and the output of the full pipeline will be Farsi script, while the internal work will be done using this romanized text, for ease of use. The romanizer will do a simple one-to-one replacement of Farsi letters with Roman letters, which will include a subset of the English alphabet, plus some additional characters for consistency with the Farsi alphabet.

3. FOMA FSTs: The main module of the lemmatizer will be a set of FSTs created in FOMA, one for each major part-of-speech. The noun FST will look for plural suffixes and indefiniteness markers, the adjective FST will look for agreement markers, and some simple parts-of-speech like particles will have FSTs that simply return the surface form unchanged.

   The verb FST will be the most complex, requiring a detection of the main verb in a verb conjugation, the separation of the main verb from its affixes and auxiliaries, and ultimately returning both the past and present roots of the verb. This will require a lexicon of verb roots, which has been taken from the Dadegan treebank's valency lexicon, and a lexicon of verb conjugations, which will be created from a reference grammar. Due to the complexity of the verb FST, this component will be developed in stages: the first stage will handle only simple single-part verbs returning just a naive stem, the second stage will return both roots of the simple verbs, and the final stage will detect the main verb in a complex predicate before

running the earlier stage FST on the main verb.

6 System Analysis on Corpus A

7 Lessons Learned and Revised Design

8 System Analysis on Corpus B

9 Final Revisions

10 Future Work

References

Amir Hossein Jadidinejad, Fariborz Mahmoudi, and Jon Dehdari. Evaluation of Perstem: A simple and efficient stemming algorithm for Persian. In Carol Peters, Giorgio Di Nunzio, Mikko Kurimo, Thomas Mandl, Djamel Mostefa, Anselmo Peñas, and Giovanna Roda, editors, *Multilingual Information Access Evaluation I. Text Retrieval Experiments*, volume 6241 of *Lecture Notes in Computer Science*, pages 98–101. Springer, Heidelberg, 2010. ISBN 978-3-642-15753-0. URL `http://dx.doi.org/10.1007/978-3-642-15754-7_11`.

Karine Megerdoomian. Unification-based persian morphology. In *Proceedings of CICLing 2000*, Centro de Investigación en Computación-IPN, Mexico, 2000.

Karine Megerdoomian. Finite-state morphological analysis of persian. In *Proceedings of the CoLing Workshop on Computational Approaches to Arabic Script-based Languages*, Geneva, Switzerland, 2004.

Mohammad Sadegh Rasooli, Amirsaeid Moloodi, Manouchehr Kouhestani, and Behrouz Minaei-Bidgoli. A syntactic valency lexicon for persian verbs: The first steps towards persian dependency treebank. In *5th Language & Technology Conference (LTC): Human Language Technologies as a Challenge for Computer Science and Linguistics*, pages 227–231, Poznán, Poland, 2011.

Benoît Sagot and Géraldine Walther. A morphological lexicon for the persian language. In *Proceedings of the 7th Language Resources and Evaluation Conference (LREC'10)*, La Valette, Malta, 2010.