11-712: NLP Lab Report

Weston Feely

April 26, 2013

Abstract

This project concerns the creation of an open-source morphological analyzer for the Farsi (Persian) language. This morphological analyzer 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, optionally with linguistic markers replacing surface morphemes. While there have been previous efforts to make morphological analyzer for the Farsi language, this project aims to create a morphological analyzer that can be easily integrated into a larger natural language processing system. The resulting morphological analyzer will be included in a Farsi NLP toolkit, which will include a Farsi text normalizer, a tokenizer/word-segmenter, a part-of-speech tagger, this morphological analyzer, 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. Persian 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, with an indefinite suffix marking indefiniteness and definiteness being unmarked. 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 letters do not join with the following letter in a given word, in which case a space or a special zero-width-non-joiner (ZWNJ) 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 freely-available Farsi stemmer mentioned above, Perstem (Jadidinejad et al., 2010), will be utilized in this project. Perstem will be useful as a tool for comparison with the morphological analyzer 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. Farsi grammatical information for the development of morphological analysis will mostly come from a Farsi reference grammar and from consulting with a native Farsi speaker.

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 morphological analyzer. The data will be split into three sections: two smaller data sets for early development, and one larger data set for the final evaluation. This dependency treebank is being used for this project because a secondary goal of this project is to integrate this morphological analyzer into a larger natural language processing system, which includes a Farsi text normalizer, a tokenizer/word segementer, a part-of-speech tagger, and a dependency parser all developed 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 orthography 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 (ZWNJ) character in between the pair. This ZWNJ character can be especially troublesome, because it's hard to see when looking at Farsi text, but can cause many issues in text processing. Additionally, using ZWNJ characters appropriately is important for producing accurate Farsi lemmas.

Additionally Farsi orthography is complicated by the use of many diacritics, which are written above or below the corresponding letter in a Farsi text. Farsi diacritics are used to mark vowels and to disambiguate terms which look identical in the abjad, but Farsi texts often omit these diacritics

or use them sporadically. Because these diacritics cannot be relied upon in our data, Farsi text normalization to remove these diacritics will be a necessary pre-processing step to morphological analysis.

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 the part-of-speech (POS) tags from the treebank will be valuable for separating the treebank's tokens into different morphological categories for the initial development process.

The Farsi affix system is mostly composed of suffixes, with a smaller number of prefixes. Affixes can be used to mark possession (in the case of possessive enclitics on nouns), definiteness, person and number, as well as verbal tense, aspect, and mood. Farsi has many native affixes, but also less common borrowed affixes from Arabic which will need to be accounted for in our morphological analysis. Additionally, Farsi has some borrowed words from languages such as Arabic which have irregular forms, such as irregular plurals. These will need to be handled separately, alongside the analysis of Farsi words and affixes.

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. This is the result of the design of the Dadegan treebank, which groups together what English treebanks might consider a phrase, like "should have been eating", into a single Farsi token. These Farsi verb tokens consist of usually the present or past root of the verb, plus one or more affixes and auxiliary verbs. Morphological analysis for Farsi verbs will require recognizing which part of a verbal conjugation is the main verb, matching of the different affixes and auxiliaries that are possible in all Farsi verb conjugations, and then the past or present root of the main verb will need to be returned as the lemma along with the morphological analysis of the affixes and auxiliaries. The full analysis of the Farsi verb tokens in the Dadegan treebank may not be possible to accomplish in the scope of this project, but future development will hopefully cover analyses for a large portion of the Dadegan verb tokens.

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. Fortunately, the Dadegan treebank's tokenization scheme separates Farsi light verbs as separate tokens from their nominal component, which will allow the morphological analysis of these two parts to be done separately using the regular morphology of each part-of-speech category.

5 Initial Design

The initial design for morphological analysis development is as follows:

Dadegan Treebank → Extract Token, Lemma, POS tag triples → Text Normalization → FOMA

→ Precision-Recall Evaluation

Here are the details for each module:

1. Dadegan Treebank
   The development and testing data for this project are Farsi words written in Farsi script, taken from the Dadegan treebank.

2. Token, Lemma, POS tag triples
   In order to expedite the analysis process, all unique triples of token, lemma, and POS tag from the treebank have been extracted and formatted into a list. This list was then randomly divided into three corpora, A, B, and C, for development and testing. The lemmas are read from these corpora into the FOMA lexicon files for each POS category, and the tokens from these corpora are the input to our FOMA analyzer during testing. The POS tags serve to help organize the analyzer-building process.

3. Text Normalization
   Because Farsi text has many variant forms, based on diacritic usage, it is necessary to normalize different orthographic variants as a pre-processing step to morphological analysis. All tokens and lemmas extracted from the Dadegan treebank have been normalized using a Farsi text normalization script created for this project, which removes all diacritics, with the exception of the madde diacritic when it occurs on the letter alef, which marks /ɑ/ to contrast with /ae/. This exception was made based on morphological analyzer performance.

4. FOMA
   The FST-based morphological analyzer for this project has been created in FOMA. There is one FOMA lexicon file for each major part-of-speech in the Dadegan treebank. The full set of part-of-speech categories is the following:

   ADJ: Adjectives
   ADR: Address terms (closed-class)
   ADV: Adverbs
   CONJ: Coordinating conjunctions (closed-class)
   IDEN: Titles (closed-class)
   N: Nouns
   PART: Particles (closed-class)
   POSNUM: Post-noun numerals
   POSTP: Postpositions (closed-class)
   PREM: Pre-modifier (closed-class)
   PRENUM: Pre-noun numerals
   PREP: Prepositions (closed-class)
   PR: Pronouns (closed-class)
   PSUS: Pseudo-sentences
   PUNC: Punctuation marks (closed-class)

SUBR: Subordinating conjunction (closed-class)
V: Verbs

Most of the POS categories from the Dadegan treebank are closed-class and only require either limited analysis, or simply returning the input tokens unchanged for these categories. These closed-class POS categories include particles, prepositions, pronouns, and punctuation marks, among others. While the numeral categories of pre-noun numerals and post-noun numerals are technically open-class, since any number can be included in these categories, the analysis for these categories will simply be returning the unchanged number as the lemma, so the numeral categories can be treated like the closed-class categories.

There are four main open-class POS categories: adjectives, adverbs, nouns, and verbs. The adjective and adverb lexicons includes morphology for comparative and superlative suffixes and derivational morphology. The noun lexicon includes analysis of plural suffixes, the indefiniteness suffix, possessive enclitics, and derivational morphology. Finally, the verb lexicon is the most complex, including negative prefixes and tense and aspect suffixes for different conjugations.

The verb lexicon will eventually perform detection of the main verb in a verb conjugation with auxiliaries, as well as the separation of the main verb from its affixes and auxiliaries, and ultimately our analyzer will return the past or present root of the verb. This requires a lexicon of verb roots, which has been taken from the Dadegan treebank's verb valency lexicon, and a full verbal morphology, which will be created in FOMA using a Farsi reference grammar. Due to the complexity of the verb FST, this component will be developed in stages: the first stage will not attempt to handle verbs at all, the second stage will be an initial attempt to analyze simple verb affixes, and the third stage will have a full analysis for single-part verbs. In future work, a later stage of this project will detect the main verb in a complex predicate before running the earlier stage FST on the main verb.

5. Evaluation
   After developing the FOMA lexicon files using the different treebank corpora, the main FOMA file will be able to be queried for analysis. The tokens from each corpus will be run through the analyzer, and the resulting analysis will be saved to file. For our evaluation metric, we've chosen precision-recall because multiple analyses can be reported for a single token. We will evaluate using only the resulting lemmas from each morphological analysis hypothesis, since gold-standard lemmas are available for each token in the treebank. This evaluation will be reported for each corpus, A, B, and C, after initial development is complete.

After laying out this initial system design in FOMA, the first step towards analysis is to implement a "guesser." For our purposes, our guesser returns the original token from the Dadegan treebank for each token, lemma, and POS tag triple. This is actually not a terrible guess, since we have many closed-class POS categories that will have 100% accuracy when simply returning the token as our lemma hypothesis, since these categories have no morphology. The results of the guesser is below. For the guesser, rather than precision-recall scores, there is simply accuracy reported, since there is a one-to-one correspondance between analyses and gold-standard lemmas for

the guesser, since we do no analysis at all.

"Guesser" Lemma Accuracy

Corpus A	0.558365758755
Corpus B	0.564931906615
Corpus C	0.558602657262

6 System Analysis on Corpus A

The first round of development of the Farsi morphological analyzer includes simple affix analysis. For nouns, this includes the plural suffix, the indefinite suffix, and the pronominal enclitics for possession. For adjectives, this includes the comparative suffix, the superlative suffix, and the same pronominal enclitics as for nouns. Verbs are not handled in this first round of development. All other POS categories have a simple lexicon with a single rule that returns the input, with no modification. Analysis is done using substitution of affixes with morphological analysis markers like "+Pl" for plural, "+3P+Sg" for third-person singular, and "+Comp" for comparative.

After making the morphological analysis rules within each lexicon file, the lemmas for corpus A were read into the lexicon files, separated for each POS category. This serves as the initial vocabulary, and ensures that it is possible to generate the lemma for each token in corpus A, provided sufficient analysis rules.

After the lexicon files have been set up, the tokens from corpus A are passed to our analyzer for the first round of testing. The analysis hypotheses for each token were saved into a results file, which was filled in with the original tokens (our "guesses") for any failed analysis. Then, a precision-recall script evaluated the results. This evaluation can be seen in the following table.

Analyzer Precision-Recall Corpus A

Precision	0.72512608895
Recall	0.641321978913
F-Score	0.680654185496

7 Lessons Learned and Revised Design

The first round of development was successful in showing an improvement over the "guesser" baseline. However, many of the guesses had to be filled in to the output from the analyzer, which indicates that many analyses couldn't reach a lemma in the vocabulary of the lexicons, due to the limited amount of morphological derivation rules in the lexicon files. Now that we have an initial analyzer, the prioritized list of improvements for the next round of development is the following:

1. An initial verb morphology for verbal affixes

2. Expanded noun and adjective morphology

3. Automatically-generated affix rules from corpus A

8 System Analysis on Corpus B

This second round of development includes simple affix analysis for verbs. The negative prefixes, the durative prefix, the subjunctive/imperative prefix, the infinitival suffix, the past tense suffixes, and the suffix conjugations for person and number are accounted for. The noun and adjective morphologies have also been expanded, to include compound suffixes for each pair of possible suffixes from the original set of suffixes in the corpus A analysis.

The precision-recall evaluation for corpus B, based on the second round of development, is in the table below.

Analyzer Precision-Recall Corpus B

Precision	0.70402994935
Recall	0.658089748868
F-Score	0.680285136717

9 Final Revisions

For the final evaluation, we did a comparison of the corpus A lemmas and tokens to automatically create morphological rules for adjectives and nouns. The lemma was removed from each token in the corpus A lemma+token pairs, and new affixes were inferred from the remaining text. Although many of the affixes from this comparison were already included in the original affix set, or in the compound affix set, several new suffixes and a few new prefixes were added to the noun and adjective lexicons. New affix rules were not added to the verb lexicon from this process, because the verb tokens were simply too complex to benefit from this comparison process; most of the remaining text were not affixes but instead the remainder of the complex predicates.

The precision-recall evaluation for corpus C, based on the final round of development, is in the table below.

Analyzer Precision-Recall Corpus C

Precision	0.606180429543
Recall	0.574973668662
F-Score	0.590164798945

10 Future Work

In future work, this analyzer will be updated to more elegantly and completely capture Farsi morphology. New versions of this morphological analyzer will be released in the coming months with these changes:

1. Suffix ordering
   In order to better generalize to new data, the compound suffixes in the noun and adjective lexicons will be split into multiple suffix categories with correct ordering.

2. Analysis options
   Since this project might be useable for different kinds of morphological analysis, a simple shell script will be created to run the Farsi analyzer, allowing multiple analysis options. For example, the English word "unbelievable" might be analyzed using full analysis (+Neg believe +Poss), text-only analysis (un+ believe +able), or lemmatization (believe).

3. Better Verbal morphology
   A more complete verbal morphology will be released in stages, to allow for better analysis of Farsi verbs. Since this is a difficult problem, as described above, this process will be done in small updates to the analyzer over time.

Finally, as stated above, the complete analyzer will be included in a more comprehensive and easier-to-use Farsi NLP toolkit, which will include the text normalizer created during this project, the tokenizer/word segmenter created for a related project, a part-of-speech tagger, this morphological analyzer, and a dependency parser trained on the same data.

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.