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 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. 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, with an indefinite marker and definite 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 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 tokenizer/word segementer/text normalizer, 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 doesn't appear when looking at Farsi text, but can cause issues in text processing. Additionally, using ZWNJ characters appropriately is important for producing accurate Farsi lemmas.

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 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), pluralilty, 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 suppletive (irregular) forms, such as irregular plurals. These will need to be handled, alongside the analysis of Farsi 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. These 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.

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.

5 Initial Design

The initial design for morphological analysis development is as follows:

Dadegan Treebank → Extracted Token, Lemma, POS triples → Lemmas for each POS → FOMA

Here are the details for each module:

1. Dadegan Treebank: The development and testing data for this project will be Farsi words written in Farsi script from the Dadegan treebank.

2. Token, Lemma, POS triples: In order to expedite the analysis process, all unique triples of token, lemma, and POS tag from the treebank will be extracted and formatted into a list. This list will be randomly divided into three corpora, for development and testing. The lemmas will be read from these corpora into the FOMA lexicon files for each POS category.

3. FOMA: The main module of the morphological analyzer will be a set of FSTs created in FOMA. There will be one lexicon for each major part-of-speech in the Dadegan treebank. The full set of part-of-speech categories is the following:

   ADJ: Adjectives
   ADV: Adverbs
   CONJ: Conjunctions (closed-class)
   IDEN
   N: Nouns
   PART: Particles (closed-class)
   POSNUM: Post-Numerals
   POSTP: Post-Positions (closed-class)
   PREM
   PRENUM: Pre-Numerals
   PREP: Prepositions (closed-class)
   PR: Pronouns (closed-class)
   PSUS
   PUNC: Punctuation (closed-class)
   SUBR
   V: Verbs

   Most of the POS categories from the Dadegan treebank are closed-class and will only require either limited analysis, or simply returning the input tokens unchanged for these categories. Among the open-class POS categories, the noun lexicon will include analysis of plural suffixes, indefiniteness markers, and possessive enclitics. The adjective and adverb lexica will include analysis of comparative and superlative suffixes.

   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 the past or present root of the verb. This will require 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 be an initial attempt to analyze simple single-part verbs, the second stage will have a full analysis for the simple verbs, and the final stage will hopefully detect the main verb in a complex predicate before running the earlier stage FST on the main verb.

   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 precision-recall will be reported for each section after initial development is complete.

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.