Feasibility of Automatically Bootstrapping a Persian WordNet

Chris Irwin Davis, Dan Moldovan
Human Language Technology Research Institute
University of Texas at Dallas, Richardson, Texas
E-mail: cid@hlt.utdallas.edu, moldovan@hlt.utdallas.edu

Abstract

In this paper we describe a proof-of-concept for the bootstrapping of a Persian WordNet. This effort was motivated by previous work done at Stanford University on bootstrapping an Arabic WordNet using a parallel corpus and an English WordNet. The principle of that work is based on the premise that paradigmatic relations are by nature deeply semantic, and as such, are likely to remain intact between languages. We performed our task on a Persian-English bilingual corpus of George Orwell's Nineteen Eighty-Four. The corpus was neither aligned nor sense tagged, so it was necessary that these were undertaken first. A combination of manual and semi-automated methods were used to tag and sentence align the corpus. Actual mapping of English word senses onto Persian was done using automated techniques. Although Persian is written in Arabic script, it is an Indo-European language, while Arabic is a Central Semitic language. Despite their linguistic differences, we endeavor to test the applicability of the Stanford strategy to our task.

1. Introduction

Modern Persian, also known as Farsi, has more than 39 million speakers[1], yet remains a language with limited NLP resources. One of the most useful tools in the NLP arsenal for any given language is a WordNet[2], but bootstrapping a new WordNet is a labor intensive process. Motivated by work on an Arabic WordNet bootstrapping system at Stanford University, the SALAAM system (Diab, 2004), we attempt to see if the same techniques are applicable to Persian.

The underlying concept of SALAAM is that "paradigmatic relations are deeply semantic in nature that they, by and large, tend to hold cross linguistically" (Diab, 2004). In other words, although there is not one-to-one correspondence between lexical entries of different languages, there does tend to be a monotonic correspondence between synsets of languages. The essential strategy is to align words with synset tags in the English half of a parallel corpus to their equivalent words in the Persian half of the corpus. Then we define entries in the Persian WordNet using the mapped information from English WorldNet. We also allow for alignment fertility which permits a single English word to map to multiple Persian words, creating a phrasal entry in the Persian WordNet.

1.1. Distinction from SALAAM

Our system used the MULTEXT-East[3] (Erjavec, 2004) English corpus of George Orwell's *Nineteen Eighty-Four* and WordNet 3.0 senses. The Persian corpus of *Nineteen Eighty-Four* is based on the MULTEXT-East framework (QasemiZadeh and Rahimi, 2006), but not officially included with that package at this time, and not yet sentence aligned with the English corpus. Performing a bilingual sentence alignment prior to sense mapping was therefore necessary. Additionally, sense-tagging is not present in the MULTEXT-East corpus, so we chose 15 word lemmas in the English text, and manually sense-tagged them for this task. These word types were chosen for frequency in the corpus to minimize data sparsity issues. Deference was also given to word types with at least moderate polysemy to adequately evaluate our system.

The SALAAM system used the SensEval 3 bilingual, word-aligned English-Arabic corpus, with 447 word types in the English half annotated with WordNet 1.7 senses. The sense-tagged English words were then mapped to their Arabic word-aligned counterparts and manually evaluated.

The SALAAM system additionally leveraged the semantic information that is embedded in the morphology of Arabic. Many words in Arabic are based upon a three letter stem called a *masdar*, which maintains a core sense context in all words in which it is present. Regular morphology rules extend both the lexical form and the semantics in predictable ways. Persian has no such morphological abstraction as the Arabic masdar, thus our system is "naïve" in that it uses no linguistic rules for learning Persian word senses.

2. Naïve Persian Bootstrapping System

Because the English-Persian bilingual corpus was not aligned, it was first necessary to perform an alignment task. This was done in two phases – a sentence alignment, then a word and phrase alignment between the aligned sentence pairs.

Sentence alignment was performed without using linguistic information. Both halves of our corpus had both paragraph and sentence tags. We performed alignment based upon matching the longest window of contiguous number of paragraphs that had the the same number of sentences. When a misalignment of sentence counts was encountered, the previous window was closed and

[1] Ethnologue 2005 estimate

[2] http://wordnet.princeton.edu

[3] http://nl.ijs.si/ME

appended to the set of aligned paragraphs. A new window was then started, and the remaining parts of the two halves shifted against each other looking for the next sequence of paragraph alignments. Once the corpus was fully processed and we had a list of aligned paragraphs, we assumed a parallel alignment of their constituent sentences. In this way we were able to identify 3260 sentence pairs to use as input to the subsequent word and phrase alignment training.

2.1. Corpus Annotation

Of these candidate sentence pairs we manually identified 15 of the most frequent English nouns and verbs to be sense tagged (Table 1). Word frequency was the primary criterion taken into account in order to minimize issues due to data sparsity, as 3260 sentence pairs is considered an extremely small data set for word alignment training. Some morphological forms were considered along with the base forms to further minimize sparsity problems.

For nouns, regular English plurals (those formed with 's') were included in addition to the singular forms. For verbs, both the past participle and third person singular were included. Adjectives and adverbs were not considered for this task.

The list of candidate words was then manually tagged with WordNet 3.0 senses in the English corpus. For this prototype, only senses with single tokens in the English corpus were considered. Some verbs were present as parts of phrasal verbs. For example, instances of "make" were found in the phrasal verb "make up" (to comprise). In these cases, the primary, non-phrasal verb sense was used.

The sense tagging was done by appending an affix to each word instance in the text, e.g. an instance of "voice" which was identified as WordNet sense number two of the noun was rendered as "`voice_n2`". This forced the following word and phrase alignment training to recognize each word sense as a unique word.

Words	WordNet		1984 Corpus	
	noun	verb	noun	verb
party	5	1	1	0
time	10	5	4	0
face	13	10	4	1
think	4	13	2	2
moment	6	0	3	0
man	11	2	2	0
eye	5	1	1	0
war	4	1	3	0
know	1	11	0	7
voice	11	2	3	0
say	1	11	0	7
mind	7	6	6	2
make	2	49	0	21
world	8	0	5	0
remember	0	8	0	2
Total	88	120	34	42

Table 1: Distinct sense counts of the fifteen manually tagged word types

2.2. Word and Phrase Alignment

Word alignment was achieved using IBM word alignment models (Brown et al., 1994). We used the GIZA++ tool set (Och and Ney, 2000) for this. GIZA++ training was seeded using bilingual word classes learned with the maximum likelihood criterion implemented with `mkcls` (Och, 1999).

IBM word alignment models account for alignment fertility (a one-to-many mapping), by assigning fertility probability to each word. Further, a mapping of 1:0 is called the null alignment, and is used for source words without an equivalent in a specific target instance.

The IBM models do not account for reverse fertility, i.e. a many-to-one mapping, therefore the word alignment training was performed in both directions. The bi-directional training output was used to generate alignment templates for each of the sentence pairs (Koehn et al., 2003). Aligned words and phrases were then extracted from the templates to create the word and phrase lookup table.

Each phrase table entry was assigned a probability based upon the frequency in the corpus of a particular aligned phrase pair. Morphological forms of the same word were considered to contribute to the same probability entry. The probability space was defined by the lexical English entry.

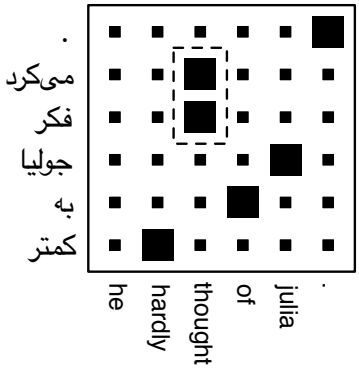

Figure 1: Example of an alignment within a phrase table entry

2.3. Identification of Persian Synsets

Employing both the phrase table and its precursor word alignments, English sense-tagged words were used to generate Persian synsets by finding their aligned equivalents. This was done using the following heuristic:

- First, prefer 1:1 mappings of English-to-Persian words in the phrase-table of the English sense-tagged words, i.e. those phrases that consist of single-word entries. For multiple candidates, prefer the most probable entry based upon frequency of occurrence in the corpus. Assign the English sense to the Persian lexical entry.
- Next, prefer 1:*n* mappings of English-to-Persian word-to-phrase entries in the phrase table. Assign the English sense to the Persian *n*-gram. Consider this a phrasal entry. For multiple candidates, again prefer the most probable entry.

Word	Noun Senses	Good	Bad	NSF	Verb Senses	Good	Bad	NSF
party	1	1						
time	4	2	2					
face	4	1		3	1		1	
think	2	1		1	2	2		
moment	3	2		1				
man	2	2						
eye	1	1						
war	3	2		1				
know					7	3	2	2
voice	3	2		1				
say					7	7		
mind	6	3	2	1	2		1	1
make					21	4	4	13
world	5	4		1				
remember					2	2		
Total	34	21	4	9	42	18	8	16
Total w/o NSF	25	21	4	–	24	18	8	–

Table 2: Training results by word

- If no single-word entry is available for the sense-tagged English word in the phrase table, consider instances where the tagged English sense exists within a phrase-table entry, and look for an unambiguous (that is, 1:1 or 1:*n*) word mapping in the supporting word alignment template data, as in Figure 1.
- Finally, if none of the previous matches can be made, mark the English sense instance as "not aligned".

These results are detailed in Table 2, and are broken down further by both noun and verb.

3. Evaluation

For each of the 76 word senses that were tagged in the corpus, we categorized the result into one of three cases. This categorization was based upon a manual evaluation by a native Persian speaker and a monolingual Persian dictionary. The possible results were:

- Good – The system chose a Persian lemma which correctly mapped to the English word sense. We categorized these as good alignments. Of particular interest were when multiple distinct Persian matches were discovered for different senses of the same English word.
- Bad – The system chose a Persian lemma which did not map to the associated English word sense. We categorized these as bad alignments.
- NSF – When the system found no corresponding Persian lemma for an English word sense, we categorized these NSF, or insufficient, data to achieve an alignment.

From this, we note the following interesting results:

time (sense n2) – The idiomatic Persian phrase "for always" was correctly identified as the equivalent for the English "all the time".

man (sense n3) – The correct synonyms for "mankind" and "human" were identified.

say – All senses aligned well, even though seven English senses mapped to five distinct Persian *n*-grams.

mind – The noun with the highest polysemy. The three senses that were categorized "good" mapped to three distinct Persian *n*-grams.

make – The word with the highest polysemy. More than half of these senses had only one or two occurrences in the corpus, and were too sparse too align. This word contributed most to the "NSF" probability space. We attribute this to the high frequency of it s mapping to the equivalent Persian word (کردن) /kardan/, which is the most common light verb component in Persian complex predicates.

remember – The partially idiomatic phrasal verb "bring memory" (خاطر آوردن) /xâter âvardan/ was identified correctly.

We discovered that out of the 25 "NSF" (no alignment) cases, 17 were associated with having only a single occurrence of the word sense in the corpus. Examining the source sentence pairs further revealed that more than half of those 17 unique instances were part of a misaligned sentence pair.

Unlike the SALAAM system, we did not consider close matches, i.e. English word senses aligned to a hypernym or hyponym in the target language. Such relations were categorized as "bad" alignments in our overall system performance. Only "good" alignments contributed to the overall performance of 51.32% correct. This compares favorably with SALAAM which scored 52.3% overall accuracy.

For the evaluation we generated two sets of performance data: one with NSF included in the probability space and the other with NSF removed. In each, we generated a positive performance percentage by considering only the number of "good" alignments within the associated probability space.

These results are summarized in Table 3.

We also see that the performance of identifying verbs was consistently inferior to that of identifying nouns. We primarily attribute this to the difficulty introduced by the multiple morphological forms of each verb due to inflection. However, we note that while our system

performed marginally better than SALAAM, the gap between noun and verb performance in our system was greater. We ascribe this larger gap to the high percentage of complex predicates in Persian. As a consequence, this lead to a greater number of 1:*n* alignments whose disambiguation introduced the higher error rate. As expected, we subsequently found a substantial number of phrasal entries among the verbs in the Persian WordNet.

	Arabic SALAAM	**Persian**	
		w/ NSF	**w/o NSF**
Noun	52.3%	61.76%	84.0%
Verb	37.7%	42.86%	75.0%
All POS	52.3%	51.32%	79.59%
Hyponym	39.96%	N/A	
Hypernym	7.8%	N/A	

Table 3: Evaluation

4. Conclusion

Although Arabic and Persian are linguistically very dissimilar, the technique used in our system resulted in a similar overall success rate to that of SALAAM for mapping English WordNet synsets to a new language. Furthermore, our system did not leverage any linguistic information as did SALAAM. We also similarly found that verbs synsets were not as successful in accurately mapping as nouns. This would suggest a *general* applicability of bootstrapping a WordNet from a word-aligned bilingual corpus with English sense tags. Recent work by Farreres et al. (2010) describes a theoretical underpinning for such an assertion.

We temper our optimism with the caveat that we evaluated only 15 word types, and our data set was extremely small. Indeed, almost a third (32.89%) of our tagged word sense candidates had insufficient data to generate a mapping. Still, some senses with only a single instance in the corpus were able to be correctly identified. While we hesitate to claim definitive results of the general applicability of this strategy to any language, overall we judged it to be effective for the automatic bootstrapping of a new Persian WordNet.

Based on our successful prototype, we are now working on the release of the first freely available Persian WordNet for the research community.

5. References

Peter F. Brown, Vincent J.Della Pietra, Stephen A. Della Pietra, Robert. L. Mercer. 1994. The Mathematics of Statistical Machine Translation: Parameter Estimation, *Computational Linguistics*, 19(2):263-311

Mona T. Diab. 2004. Feasibility of Bootstrapping an Arabic Wordnet Leveraging Parallel Corpora and an English Wordnet. In *Proceedings of the Arabic Language Technologies and Resources, NEMLAR*.

Tomaš Erjavec. 2004. MULTEXT-East Version 3: Multilingual Morphosyntactic Specifications, Lexicons and Corpora. In *Procedings of the Fourth International Language Resources and Evaluation Conference, LREC'04*.

Javier Farreres, Karina Gibert, Horacio Rodríguez, Charnyote Pluempitiwiriyawej. 2010. Inference of lexical ontologies. The LeOnI methodology. *Artificial Intelligence*, Vol. 174, No. 1. (25 January 2010), pages 1-19.

Philip Koehn, Franz Josef Och, and Daniel Marcu. 2003. Statistical Phrase-Based Translation. In *NAACL/HLT 2003, Proceedings of the Human Language Technology and North American Association for Computational Linguistics Conference*.

Franz Josef Och. 1999. An Efficient Method for Determining Bilingual Word Classes. *Ninth Conference of the European Chapter of the Association for Computational Linguistics*, pages 71-76, EACL'99.

Franz Josef Och and Hermann Ney. 2000. Improved Statistical Alignment Models. In *Proceedings of ACL 2000*, pages 440-447.

Marian Olteanu, Chris Irwin Davis, Ionut Volosen, and Dan Moldovan. 2006. Phramer - An Open Source Statistical Phrase-Based Translator. In *Proceedings on the Workshop on Statistical Machine Translation*, pages 146-149, New York City, Association for Computational Linguistics.

Behrang QasemiZadeh and Saeed Rahimi. 2006. Persian in MULTEXT-East Framework In *Advances in natural language processing*, volume 4139 of 5th International Conference on NLP, FinTAL, pages 541-551.