Translating from Universal Networking Language into Persian Language

Mohammad Hossein Ariana
Department of Computer Engineering
Dehdasht Branch, Islamic Azad University
Dehdasht, Iran

Hassan Rashidi[2]
Department of Mathematics and Computer Science
AllamehTabataba'i University
Tehran, Iran

***Abstract*- In this paper, we have discussed the machine translation using Interlingua. The intermediate language used is Universal Networking Language which has been developed by the University of United Nations in Tokyo. We have provided software and a set of rules to convert from this Interlingua into Persian language which we'll discuss here. An important facet of the work is preparing a categorization of Persian language sentences. Implementation of this translator involves creating a special type of Persian-UNL dictionary of words, providing grammatical rules of Persian Language and finally suggesting a routine for converting UNL sentences into their Persian equivalent. We show that using this method, translation from other languages into Persian can be improved dramatically.**

Keywords—Deconverter, Universal Networking Language, UNL, Persian, Machine Translation, Generation Rules

I. INTRODUCTION

In recent years, with the increasing use of the Internet, different machine translation methods have been applied to the web environment. Increase in amount of information available on the web from one side and increase in number of the Internet users from the other side, has caused a justifiable interest for using machine translation.

Machine translation methods are generally divided into three categories. The first method is called Transfer method in which the source language text is analyzed in some depth and then is translated into target language sentences using grammatical rules of target language. The second method is called the Interlingua approach, in which translation is performed using an intermediate language, which is a representation of concepts independent of any language and the translation process can be performed independent of source or target language. The third approach is statistical method, in which frequency of each word is determined and when translating, a word is chosen according to its probability of occurrence [1].

In this paper, we have used the second method, i.e. the Interlingua approach. The Interlingua used is Universal Networking Language which has been developed in United Nations University in Tokyo to overcome the language barrier on the web. We describe translation from this Interlingua into Persian Language, i.e. the Deconversion process. When using an Interlingua, the necessary rules are designed using the meanings of the words and the sentences. From the other side, in most of natural languages, the meaning of each word or phrase depends on its grammatical rule. Persian is an ancient language and during its existence, has experienced a lot of changes. Hence, the rules provided should be complete enough to deal with different types of sentences[2].

The rest of the paper is organized as follows: Section 2 discusses literature review. A summary of Universal Networking Language is given in Section 3. Design of Persian-UNL word dictionary and translation rules is presented in Section 4. Section 5 discusses the implementation of the system. Conclusion and future works are given in Section 6.

II. LITERATURE REVIEW

A lot of work has been done in translating from/to Persian language, but most of them used the transfer method described above, i.e. translating from a specific language into Persian. For example, in [3] authors have proposed a method to translate from Persian into English language which recognizes and converts Persian sentences according to a structure.

Another example is a huge project named "Shiraz Project" [2] which is used to translate Persian into English language and the researchers have done a lot of job on Persian language structure and grammar. They have also developed an application program for their method. This project uses a chart based approach which is a relatively new method, but the drawback is that in this work, the Persian words should be entered using English alphabet, a phenomenon known as Finglish[4].

A semantic parser has been proposed in [5] which enable us to convert the Persian sentences into an interlingua. This work is one of the early papers in which the Persian language is considered as language with free words, i.e. the place of sentence words like the subject, the object and the verb could be changed with the meaning of the sentence remains unchanged. Yet, in this work, they do not make any use of semantic structure of the sentences which makes understanding the resulted sentences difficult.

In [6] the authors have done the word analysis, the structure analysis and other types of analysis for Persian language.They have compared their work with the previous ones like the Dena system.

III. UNIVERSAL NETWORKING LANGUAGE

Universal Networking Language is a semantic representation which has enough capability to express content in natural languages. This language, which is called UNL, has been proposed by Dr. Hiroshi Uchida in University of Advanced Studies of United Nations in Tokyo. It is best suited for use as an intermediate language in translation, but is applicable in text summarization and information retrieval [7].

A. UNL components

UNL is a language for computers and has got all the necessary components of a natural language required for representation of information and knowledge. The language is composed of words called Universal Words which describe the concepts. These words or *UW*s connect to other UWs to create UNL expressions of sentences. These connections which are called *relations* specify the role of each word in the sentence. Special meanings, depending on author's point of view, can be expressed using the related *attributes* [8].

B. Relations

Relations are a series of semantic concepts which can be used in natural languages. For example, the concept of agent or reason of an event is one of such concepts. Current UNL specifications consist of 41 conceptual relations. Each relation's name is a string of three or less letters of the English alphabet.

Choosing the right conceptual relations with the correct UWs, allows us to express the propositional contents of nearly any natural language sentence. For example, in the sentence "The boy eats an apple in the kitchen", there exists a main predicate "eats" with three additional parameters which two of them are of reasoning relation type; "the boy" is the one who does the predicate "eats" and "an apple" is the object. Also, there exists a place relation: "the kitchen" is physical place where the given work is done.

C.Attributes

The attributes represent the grammatical properties of the words. They show what is said from the speaker's point of view. We also use attributes for expressing the scope of concepts; i.e. being typical or general.

Conceptual relations and UWs are used to describe the objectivity of sentences. Attributes of UWs enrich this description with more information about how the speaker views these state of affairs and his attitudes toward them.

For example, the corresponding UW of play is "play (icl>do)". If the word "play" is in the past form in the sentence an attribute @past is tagged with "play (icl>do)".

D. Universal Words (UWs)

UWs are the words which constitute the UNL vocabulary. So, UNL words are named UW. The designers of UNL have proposed that developers use English language words as UWs which are restricted using some semantic constraints.A UW is made up of a character string (an English-language word) followed by a list of constraints.

<UW> ::= <headword> [<constraint list>]

Headword of a UW is an English expression that is interpreted as a label for a set of concepts: the set made up of all concepts that may correspond to that in English. Constraint list restrict the concept of a UW to a subset or to a specific concept included within the Basic UW [9].

IV. ARCHITECTURE OF UNL-PERSIAN DECONVERTER

To convert from the UNL into other languages, UNDL foundation has designed a software program called Deco which, using a set of rules for each natural language, can convert texts to any of those languages [10]. For the sake of simplicity of the grammatical rules and optimization, we have designed our own Deco software.

A. Implementation of Persian Dictionary

Our specific dictionary consists of a set of Persian words accompanied by their equivalent meanings in the UNL language. We have collected a set of frequently used Persian words and stored them with their English meanings in our dictionary.

The entries of our dictionary have the following format. Meaning of each parameteris described fully in [11].

[HW]{ID}READING"UW"(ATTR ,...)<FLG ,FRE ,PRI>;

B. Categorization of words

First, we present our categorization of Persian words and introduce the attributes we have assigned to each category. We'll use these attributes in the next stage.

We have classified the words into six categories of noun, verb, adjective, adverb, pronoun and preposition and represent those using abbreviations of N, V, Adj, Adv, Prn and Prep, respectively.

In the case of Nouns, we have specified four categories of plural/singular, animate/inanimate, proper name and countable/non-countable names and assigned abbreviations of SIN/PLU, LIV/NOTLIV, SP and CNT/NCNT to them, respectively.

Regarding verbs, we have specified six categories of tense, person, transitive and intransitive, active or passive, noun-clause and verb-clause statements. Also, the tenses of these verbs have been categorized into three groups of past, present and future which are assigned the same group names as their tenses. Person of each verb has been specified according to number and count into six groups of 1SG, 2SG, 3SG, 1PG, 2PG and 3PG. For transitive and intransitive verbs, we have assigned the abbreviations of TRANS and INTRANS. Singular and plural verbs are shown using VS and VP, respectively. Abbreviation of active and passive verbs is VACT and VPAS, respectively. Noun-clause and verb-clause verbs are indicated using VDO and VNDO.Adjectives are represented using ADJ abbreviation.

Adverbs are shown using ADV abbreviation and adverbs of time, place and quality are shown using TIM, PLC and STAT abbreviations, respectively.

Pronouns are divided into four categories of personal, possessive, objective and determiner ones for which abbreviations of PPRON, POSS, OBPRON and POINT are used respectively.

Finally, particles are classified as conjunctions, prepositions, object sign, coordinators and interjections are shown using LINK, EXT, OBJP, CONJ and ADDR abbreviations.

Here are some examples from the Persian language:

[ما] {} "we" (1PG,PRON,PPRON)

(phonetics: *mʌ*)

For word “*ما*” in Persian we have chosen universal word “we” and attributes:1PG shows that this word is first person and plural, PRON shows that it is a pronoun and PPRON shows that it is a personal pronoun.[بود]{}"is"(V,1SG,PAST,INTRANS,SIN,VNDO)

(phonetics: *bu:d*)

It means that instead of the Persian word of “*بود*” we have chosen universal word “is”. Companion attributes indicate that this word is a verb, in singular form and first person, in past tense, intransitive, singular and finally a noun-clause kind of a verb.

After preparing a list of frequently used words, we feed these entries into our dictionary system and use them as our knowledge base in the next phases.

C.Preparing generation rules

At First, we present the categorization of sentences in Persian. This categorization is useful in the sense that when generating Persian sentences, we know how to create syntactically correct sentences.

D.Categorization of Persian sentences

Generally, Persian sentences are either nominal or verbal [12].Every sentence with a linking verb is nominal, otherwise it is verbal. Linking verbs in Persian are “است” (is), “بود” (was), “شد” (became), “گشت” (become), and“گردید” (become).

Rule 1: each sentence whose verb is one of the linking verbs (است, بود, شد, گشت, گردید), is a nominal, otherwise it is verbal.

Now, using rule 1 we present following rules:

Rule 2: if a sentence is nominal, then it has at least two parts of speech; the complement and the verb ones.

Rule 3: if a sentence is verbal, then it must have the subject and the verb parts of speech.

After this categorization, we present yet another categorization to help facilitate determining words part of speech in sentences. Each sentence is made up of two parts: subject and predicate such that subject is subject of both nominal and verbal sentences and predicate is equivalent to verb-phrase and probable noun-phrases accompanying it in nominal and verbal sentences.Now, we categorize the Persian sentences based on the number of the noun-phrases using a set of rules.

Rule 4: A sentence is called a two-part compound sentence if its subject is a noun-phrase and its predicate is a verb-phrase.

Rule 5: A sentence is called a three-part compound sentence if it is comprised of two noun-phrases and a verb-phrase.

The subject is always composed of a single noun phrase. Actually, in three-part compound sentences a noun—phrase is added to the predicate, which may take one of three parts of speech, depending on the sentence being nominal or verbal.

1. Object: in the sentence “کلاغ صابون را خورد” (the crow ate the soap) , “صابون” (the soap) is the object.
2. Subject complement: in the sentence “هوا سرد شد” (it got cold), “سرد” (cold) is the subject complement.
3. Complement: in the sentence “مرد با اتوبوس آمد” (the man came by bus), “اتوبوس” (the bus) is the complement.

Using these descriptions, we can determine predicate’s noun-phrase part of speech using two following rules:

Rule 6: If a sentence is a three-part compound and nominal, then the noun-phrase of the predicate is a subject complement.

For example in the sentence “““هوای گرم این شهر غیرقابل تحمل است (This city’s hot weather is intolerable), “هوای گرم این شهر” (this city’s hot weather) is the subject and “ غیرقابل تحمل است” (is intolerable) is the predicate. So, using rule 6, we know that “غیرقابل تحمل” (intolerable)is the subject complement.

Rule 7: If a sentence is a three-part compound and verbal and

a) If the noun-phrase of the predicate begins with a preposition, then this noun-phrase is a complement.

b) Otherwise, it is an object.

The last type of sentences in Persian, are four-part compound sentences. The associated rules are given afterwards.

Rule 8: A sentence is a four-part compound sentence if and if it is comprised of three noun-phrases and a verb-phrase.

In these sentences, the important thing is to determine the part of speech of the noun-phrases of the predicate. These two noun-phrases may appear in four forms:

1- object + complement: in sentence” آن مرد فرزانه، مطلب مهمی را به من یاد داد” (That wise man taught me an important thing), “مطلب مهمی” (an important thing) is the object and “من” (me)is the complement.

2- object + subject complement: in sentence” آن پسر پدر خود را قهرمان میدانست” (The boy considered his father a hero), “پدر خود” (his father) is the object and “قهرمان” (a hero) is the subject complement.

3- complement + subject complement: in sentence” مردم روستا به او دکتر میگفتند.” (The villagers called him “the doctor”), “او” (him) is the complement (in English is the direct object) and “دکتر” (the doctor)is the subject complement (of course in English language this is an object complement).

4- object + object: in sentence” دانش آموزان معلم خود را کتاب خوبی هدیه دادند” (Students rewarded their teacher a good book), “معلم خود” (their teacher) is an object and “ کتاب خوبی” (a good book)is another object.

Actually, the verbs presented in rule 1, always take part in nominal sentences and these sentences always have three parts of speech; namely the subject, the subject complement and the linking verb. Sometimes, these verbs are called pure nominal sentences.

There exists yet another group of verbs in Persian which may appear in both nominal and verbal sentences. These verbs are called the ambitransitive verbs. Some more examples:

- In the sentence “ علی خانواده اش را شاد (گردانید/ کرد/ ساخت/ نمود)” (Ali made his family happy), “شاد” (happy) is an object complement.
- In the sentence “من او را عاقل (پنداشتم/ دیدم/ دانستم/ یافتم)” (I regarded him a wise man), “عاقل” (wise) is an object complement.

Verbs like above verbs which are placed in the parentheses, are ambitransitive verbs. If a verb is ambitransitive, then the sentence containing it is also ambitransitive [13]. Now, we can state the rules of the four-parted compound sentences.

Rule 9: if a sentence is a four-parted one and

1) One of noun-phrases of the predicate begins with a preposition, that noun-phrase is a complement and the other noun-phrase is
   a. A subject complement, if the sentence is ambitransitive.
   b. An object, otherwise.
2) None of the noun-phrases begins with a preposition
   - If the sentence is not ambitransitive, both of the noun-phrases are objects.
   - Otherwise the noun-phrase with the direct object marker is the object and the other is a subject complement.

D. Rule extraction

Here we mention the main operations in translation in which we translate UNL expressions into their Persian language equivalent sentences.

D1: Two-parted verbal sentences

Two-parted sentences are made up of two parts; the subject and the verb. For example in the sentence “ علی آمد” (Ali came), ‘Ali’ is the subject and “آمد”(came) is a verb.. So, the UNL equivalent of this sentence is as follows:

Agt(Ali, Amad)

So, we use the following rule:

Rule 10: if a sentence contains a single UNL relation Agt(verb, subject), then it is equivalent to a two-parted verbal sentence.

The appropriate subject and the verb should be obtained using the existing attributes of the words in the relation. So, the general form of produced sentence is:

(Appropriate subject) (Appropriate verb)

D2: Three-parted nominal sentences

Three-parted nominal sentences are made up of three parts; subject, subject complement and linking verb. For example, “ کتاب سنگین است”(the book is heavy) is a three-parted sentence in which “کتاب”(the book) is the subject, “سنگین”(heavy) is the subject complement and “است”(is) is the verb. This sentence assigns the state of being heavy to the book. In the UNL equivalent of this sentence, we express the relation between the subject and the subject complement. So, the UNL translation of this sentence is as follows:

Aoj(“sangin”,’ketab’)

Rule 11: if a UNL expression contains only a single relation as in the following form, its equivalent Persian sentence is of three-parted nominal type.

Aoj(subject complement, subject)

So, the general form of generated sentence is as follows:

(Subject) (Subject complement) (Linking verb)

D3: Three-parted verbal sentences

We know from rule 7 that three parts of a three-parted verbal sentence are a subject, a verb and either an object or a complement. So, three-parted verbal sentences are of two types:

a) Three-parted verbal sentences with complement

These groups of sentences are composed of prepositions and complement, in addition to subjects and verbs. So, in addition to relations between the subject and the verb in the sentence, we must determine complement’s role in the sentence. The meaning of the sentence or the manner in which an event has occurred depends on the preposition. So, for each and every complement in these sentences, we should find a different UNL relation.

So we express rule 3 for translation of three-parted verbal sentences with a complement.

Rule 12: If a UNL expression contains two relations in the following form, then its Persian equivalent sentence is a three-parted verbal sentence with a complement.

Agt(verb, subject)

Relation(verb, complement)

The second UNL relation shows a relationship between the complement and the verb. Different prepositions cause the sentence mean differently. So, the general form of generated sentence in Persian is:

(Subject) Preposition (complement) (verb)

We have found the appropriate UNL relations for all the prepositions currently being used in Persian and used them in implementation of our translation software.

b) Three-parted verbal sentences with object

Another set of three-parted verbal sentences contains an object. Usually, a direct object marker “ra” accompanies the object. So, a sentence like “علی کتاب را آورد”(Ali brought the book) is a sentence with an object. So, if translated into UNL, we would have:

Agt(avard , Ali) [Agt(brought , Ali)]

obj(avard , ketab) [obj(brought , book)]

Rule 13: If a UNL expression is in the form of two following relations, its Persian equivalent is a verbal sentence with an object:

Agt(Verb, Subject)

Obj(Verb, Object)

So, the general form of this sentence is:

(Subject) (Object) “ra” (verb)

D4: Four–Parted Compound Sentences

Four-Parted sentences are composed of two noun-phrases, in addition to two main parts of subject and predicate. These two noun-phrases may appear in four different forms. Having this description in mind and also using rule9, we can find UNL relations equivalent to Persian language sentences.

Rule 13: If a UNL expression is in the form of two following relations, then its Persian equivalent sentence is a four-parted sentence and one of its components is a complement.

Aoj(subject complement, subject)

Relation(subject complement,complement)

So, the generated sentence in Persian language has the following format:

(Subject) (Appropriate preposition) (Complement) (Subject complement) (Linking verb).

Rule 15: If a UNL expression is composed of following UNL relations, then its Persian equivalent sentence is a four-parted sentence with a complement and an object.

Agt(Verb, subject)
Obj(Verb, object)
Relation(Object , complement)

So, the Persian sentence is in the following form:

(subject) (appropriate preposition) (complement) (object) “ra” (verb).

Or

(subject) (object) “ra” (appropriate preposition) (complement) (verb).

Rule 16: If a UNL expression is expressed in the following UNL relations, then its Persian equivalent sentence is a four-parted sentence with two objects:

Agt(verb , subject)
Obj(verb , first object)
Ben(verb , second object)

So, the general form of these sentences is:

(subject) (object #1) (ra) (object #2) (verb)

Rule 17: If a UNL expression is composed of the following UNL relations, then its Persian equivalent sentence is a four-parted sentence with a subject complement and an object. The word with the object marker “ra” is the object:

Aoj(verb , subject)
ben(verb , subject complement)
obj(subject complement , object)

The general form of the equivalent Persian sentences is as follows:

(subject) (object) “ra” (subject complement) (verb)

V- IMPLEMENTATION

As shown in the previous section, we can design a software application which, with some manipulation, can convert UNL phrases to their Persian equivalents. We perform this job in a sentence based method. First, we get UNL text from the input. Each sentence in UNL is placed between {org} and {/org} tags. So, one can easily separatesentences from eachother. We have used Visual Basic programming language for implementing the application. There exist some cases:

If a sentence is made up of only a single UNL relation in the form of “Agt”, then itmust be a two-parted verbal sentence and we can generate a correct sentence from it using UNL dictionary.

If a sentence is made up of a single UNL relation in the form of Aoj, then it must be a three-parted nominal sentence and we should find three parts; subject, subject complement and appropriate linking verb using the UNL dictionary.

If a sentence contains two UNL relations and one of them is “Agt”, then the sentence is a three-parted one with a complement and we should find three parts; subject, complement and the verb using the UNL dictionary. In addition, the preposition should be chosen according to the second UNL relation.

If a sentence is made up of two UNL relations of “Agt” and “Obj”, then this sentence is a three-parted compound sentence with a subject and we should find three parts of sentence; namely the subject, the object and the verb using the UNL dictionary and place the object marker “ra” after the object.

If a sentence contains two UNL relations and one of them is "Aoj”, then the sentence is a four-parted nominal sentence and we should find four parts of it; subject, complement, subject complement and linking verb using the UNL dictionary. In addition, the preposition should be chosen according to the second UNL relation.

If a sentence is made up of three UNL relations, two of them being “Agt” and “Obj”, then it is a four-parted compound sentence with a complement and object and we should find four parts of the sentence; the subject, the complement, the object and the verb using the dictionary. The preposition placed before the complement should be determined using the third UNL relation. In addition, the object marker “ra” should be placed after the object.

If a sentence is made up of three UNL relations; “Agt”, “Obj” and “Ben”, then this sentence is a four-parted compound sentence with two objects and we should find four parts; the subject, the first object, the second object and the verb using the UNL dictionary. In addition, the object marker “ra” should be placed after the first object.

If a sentence contains three UNL relations, one of them being “Aoj”, then the sentence is a four-parted compound with an object and subject complement and we should find the four parts, i.e. the subject, the object, the subject complement and the linking verb using the UNL dictionary. In addition, the object marker “ra” should be placed after the object.

VI. CONCLUSIONS AND FUTURE WORK

Nearly all available Persian translators use direct transfer methods, meaning that for translating from/to any other language into Persian, the specific rules for that language should be devised. However, using the proposed method in this paper, one can translate texts from UNL language into Persian. Actually, in direct transfer, to translate between N languages, we need N(N-1) different translation modules, but in this method we only need 2N modules.

In the future, we can work on translation for specific purposes, e.g. computer or literature and also conversational texts. We can do this job by increasing the number of dictionary entries and the number of attributes used for describing words.

REFERENCES

[1] J. R. Hobbs and Chair, ”Machine Translation” , Blackwells-ncc, London, 1993.

[2] J. W. Amtrup, H. Mansouri Rad, K. Megerdoomian and R.Zajac ,“Persian-English Machine Translation: An Overview of the Shiraz Project” , Memoranda in Computer and Cognitive Science MCCS-00-319, April 2000.

[3] S.Rezaei and M.Fahimi, *“Toward a Machine Translation System for PersianLanguage”*, 1994.

[4] A. Alipour, A. Rr. Aghayoosefi and N. AbaszadeAghdam, “the effect of reading Persian text written with English letters (Finglish) on

salivary cortisol in high school female students" , journal of school psychology summer 2013 , volume 3 , number 2 (2); pp. 124-135.

[5] M.Raeis_Ghasem, *"Natural Language Processing and Processing of PersianLanguage"*, Master's thesis, Sharif University of Technology, Tehran, 1991.

[6] (In Persian) A. S.Shahabi and A.Sarraf_Zadeh, "Machine Translation", The Linguistics Magazine, Persian Language Expansion Group, 2007.

[7] H. Uchida, M. Zhu, "UNL2005 from Language Infrastructure toward Knowledge Infrastructure", 2005.

[8] UNDL Center, "The Universal Network Language (UNL) Specification", Version 3.3, UNDL Foundation, 2004.

[9] I. Boguslavsky, J. Cardeñosa, C. Gallardo and L. Iraola, "The UNL initiative, an overview", Lecture Notes in Computer Science, vol. 3406, pp. 370-378, 2005.

[10] S. S. Spall, G. Gupta and Mr. P. Bhatia, "Architecture of UNL Punjabi Deconverter", Proceedings of COIT 2007, national conference on challenges& opportunities in Information Technology, 2007.

[11] UNL Center.(2000), Deconverter Specification, UNDL Foundation.

[12] (In Persian) A. Gharib, M. Bahar, B. Forouzanfar, J. Homayi and R. Yasemi"Persian Language Grammar (five teachers)",Naahidpunlications, Tehran, 2001.

[13] (In Persian) T. V. Kamyaar and G. R. Omrani, "Persian Language 1 (High School Book)", Education Ministry Publications, Tehran, Iran, 2005.