An Application of Lexicalized Grammars in English-Persian Translation

Heshaam Feili and **Gholamreza Ghassem-Sani**[1]

Abstract. Increasing the domain of locality by using Tree Adjoining Grammars (TAG) caused some applications, such as machine translation, to employ it for the disambiguation process. Successful experiments of employing TAG in French-English and Korean-English machine translation encouraged us to use it for another language pairs with very divergent properties, Persian and English. Using Synchronous TAG (S-TAG) for this pair of languages can benefit from syntactic and semantic features for transferring the source into the target language. Here, we report our experiments in translating English into Persian. Also, we present a model for lexical selection disambiguation based on the decision trees notion. An automatic learning method of the required decision trees from a sample data set is introduced, too.

1 INTRODUCTION

Tree adjoining grammars (TAGs) have several unique properties that make them suitable to be used by applications such as semantic interpretation and automatic translation of natural language [2, 8, 14]. This type of grammars is related to a class of grammars named Mildly Context-Sensitive Grammars (MCSGs), which is placed between context-free and context-sensitive grammars with respect to their generating power [5].

Tree Adjoining Grammars are an extension of context free grammars (CFGs), which use trees instead of productions as a primary representing structure. Formally, a TAG is a 5-tuple (V_N, V_T, S, I, A), where V_N is a finite set of non-terminal symbols, V_T is a finite set of terminal symbols, S is the axiom of the grammar, I is a finite set of *initial trees*, and A is a finite set of *auxiliary trees*. The union of I and A is the set of *elementary trees*. Internal nodes are labeled by non-terminals, and leaf nodes by terminals or empty string, except for exactly one leaf node per each auxiliary tree (called the *foot node*) that is labeled by the same non-terminal used as the label of its root node. New trees are derived by *substitution* or *adjoining* actions.

Substitution of a node labeled by A, in an elementary tree T by another tree T' rooted by label A, is performed by replacing the desired node by the whole tree T'. Let T be a tree containing a node labeled by A, and let T' be an auxiliary tree with both root and foot node labeled by A. Then, adjoining of T' into T is obtained by excising the sub-tree of T, which has a node labeled by A (i.e., called *adjunction node*), and then attaching T' to that node, and the excised sub-tree to the foot of T' [6]. Adjunction nodes are labeled by a symbol (*) and substitution nodes are labeled by a symbol (↓).

A Lexicalized tree adjoining grammar (LTAG) is a version of TAG that is lexicalized by a lexical item. In other words, in this type of grammars, every elementary tree is associated with a lexical item. The leaf node associated with a lexical item is named the *anchor* node. Anchor nodes are usually labeled by a symbol (◊).

The use of TAG for automatic translation of natural languages has led to a new concept named synchronous tree adjoining grammar (S-TAG). The use of S-TAG for machine translation was first described by Abeille et al. [1], and since then several experiments have been reported, most of which adopted the XTAG system [17]. Abeille et al. noted that traditionally difficult problems mentioned by Dorr [7], such as structural, lexical, conflation, and thematic divergences are not regarded as problems for an S-TAG based approach [10].

S-TAG is defined as two related TAGs for the source and target languages. Any sentence in the source language with its structure, which is interpreted in the TAG formalism, is related to its associated structure, again in the TAG formalism [14].

2 PERSIAN LANGUAGE

Persian, also known as Farsi, is the official language of Iran and Tajikistan, and one of the two main languages used in Afghanistan. This language has been influenced by local environments such as Arabic language (in Iran) and Russian language. Here, we use the Persian that is the official language of Iran.

Arabic language has heavily influenced Persian, but has not changed its structure. In other words, Persian has only borrowed a large number of lexical words from Arabic. Therefore, in spite of this influence, it does not affect the syntactic and morphological forms of Persian [9].

Persian is a language with SOV form with a large potential to be free-word-order, especially in proposition adjunction and complements. For example, adverbs could be placed at the beginning, at the end, or in the middle of sentences, and this does not often change the meaning. This flexibility in word ordering is usually useful in language generation.

Written style of Persian is right to left and it uses Arabic alphabet[2]. Vowels generally known as short vowels (a, e, o) are usually not written. This causes some ambiguities in pronunciation of words in Persian.

[1] Computer Engineering Department, Sharif University of Technology, Tehran, Iran, email: {hfaili@mehr.sharif.edu, sani@sharif.edu}.

[2] By Persian, we mean the language that used in Iran. In Tajikistan, Persian is written by English letters.

3 TRANSLATING ENGLISH INTO PERSIAN

English and Persian have a wide range of differences, in both structural and lexical aspects. Unlike strict SVO word order of English, Persian uses a SOV pattern with relatively free word order.

Morphological analysis of Persian differs from English in various ways. Persian morphology is an affixal system consisting of mainly suffixes and a few prefixes. There are a relatively small number of affixes in the language that obey a regular morphotactic order. These affixes are joined to the root form of words in order to produce the correct form.

S-TAG seems to be an appropriate tool to overcome these discrepancies. We've developed an S-TAG grammar for English-Persian language pair based on the XTAG project [17]. A corpus set containing 860 sentences shorter than 16 words collected from computer related articles and a set of 2136 English words were used.

We adopted an idea similar to that of an English-Korean translator [8], in which the association between peer grammars was divided into three different phases, namely: *tree transfer, lexical transfer, and feature transfer*. Figure 1, shows the process of translation using this approach. In the following sections we give a detailed account of each module.

3.1 Parsing Phase

The first phase of the translation process is *Parsing*. In this phase, each input sentence is analyzed and its structural information is extracted. Using a TAG to model the source language, and by having a parsing algorithm based on the TAG formalism, such as one described by Van Noord [16], a *derivation tree* is generated as a result of parsing each input sentence (here in English). Derivation tree is a tree that records the history of composition of the elementary trees associated with the lexical items in the sentences [17]. *Derived tree* is the syntactic structure of the sentence, which can be built by using derivation tree.

3.2 Transfer Phase

This phase, which comprises three different stages, is used to transfer an English derivation tree into the corresponding Persian derivation tree.

3.2.1 Tree Transfer Model

The basic idea of the transfer module relies on the derivation trees that are transferred from one language to another. By using the tree transfer routine, which contains all the node-to-node correspondence between elementary trees of the S-TAG, target derivation trees are built. This transfer is only a structural transfer, by which the correct related structure of Persian is generated.

In Persian, tree transfer sometimes faces some ambiguities. For example, the following English elementary tree, which handles the sentences containing *sentential complement with a noun phrase*, may be transferred into two different Persian elementary trees, based on the mode of the main verb of the embedded sentence. Sentences (1) and (2) below can be parsed by this tree[3] (The transliteration used is the same as what was described in [3]).

(1) *Srini begged Mark to increase his disk quota*
Srini az Mark barae afzayesh zarfiat disk-ash darkhast-kard
Srini Mark to increase quota disk-his begged
(2) *Beth told Jim that it was his turn*
Beth be Jim goft keh an noobat-ash bood
Beth Jim told that it turn -his was

In the first sentence, the main verb of the embedded sentence "*increase his disk quota*" is in its infinitive form, whereas the second example is in *declarative* form. Figure 2(a) shows the elementary tree for declarative sentential complement with NP, while figures 2(b, c), show the associated Persian elementary trees. In the first case, where the embedded sentence has the main verb with an *infinitive* form, the sentence appears before the main verb, while in the other one the embedded sentence conjuncts to the end of the main sentence.

This example shows a one-to-many relationship between English and Persian elementary trees. Some constraints are associated with every relation, and are used during transferring the tree structure of the source into the target language. These constraints deal only with the structural information, which is available during the transfer phase.

There may exist some relations that map many to one English-Persian elementary trees. In this case, a whole derivation tree may be transferred into an elementary tree from English to Persian. However, we haven't faced these phenomena in the sample data set that has been used.

The node correspondence between peer elementary trees (source and target languages) may have any sort of relationships: one-to-one, one-to-many, many-to-one, and many-to-many. The node correspondence is used in two cases:

- The correspondence between the anchor nodes of elementary trees, used in the lexical transfer, where the source word is transferred into the target word.
- The correspondence between other nodes of elementary trees, used during manipulation of the actions on trees (*substitution* or *adjunction*). Any action that is applied to a node in the source tree will be also applied to the corresponding node of the target tree.

3.2.2 Lexical Transfer Model

By transferring a derivation tree, the derived tree can be easily computed, and the syntactic structure of the target sentence is built. The next phase is to transfer the leaves of the derived tree, where the lexical is held[4]. This phase, which is named *lexical transfer,* is a very important but unfortunately ambiguous phase. Lexical selection becomes more difficult when it deals with some words that have more translation candidates. For example, word "*who*" can be translated into several different Persian words such as "*che kasi*" (in interrogative sentences), "*keh*" (in relative clauses), etc. This process may be even more difficult when it comes to deal with some prepositions such as "*on*", "*to*", or some specific verbs such as "*get*" and "*make*".

[3] The elementary tree and examples are borrowed from [15].

[4] Notice that the leaves of the derived trees correspond to the anchor nodes of derivation trees.

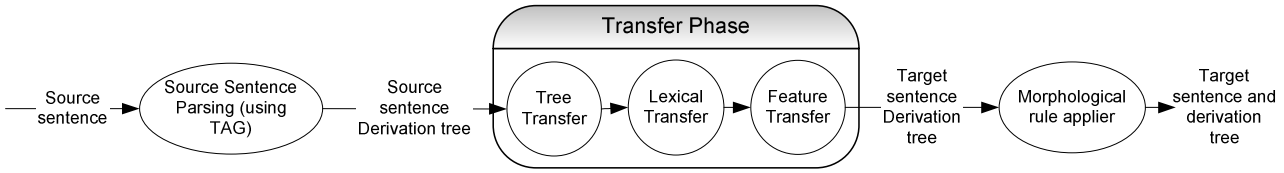

Figure 1. The process of translation using S-TAG

In general, word sense disambiguation is one of the main obstacles in the way of the lexical transfer phase [15]. Here, we introduce a new method to tackle ambiguities that occur during this phase. To that end, we build a decision tree for every ambiguous word. The decisions are made by using some attributes that are incorporated into the decision tree. These attributes are language dependent, which are instantiated in a feature-based lexicalized tree adjoining grammar (FB-LTAG) framework [17]. The attributes are defined as features in the TAG model and get their values by a unification process performed during the parsing process. The defined attributes are features that contain both syntactic and semantic information of the sentences. The main attributes used are:

- The *Part-of-speech* of the anchor node of elementary trees, which is the most important attribute to be used during decision-making. This attribute reflects the syntactic information of the sentence. For example, knowing the part-of-speech of the word "*can*" leads us to its specific meaning.
- The *family tree*[5] of the used elementary trees, which is also an important attribute to be used during decision-making, and reflects some semantic properties of the input sentence. For example, there is a family tree named "*ergative*" that refers to the verbs such as "*melt*", for which the subject plays the role of the object. So, in this family, the translated sentence should be in the passive form [17].
- The *specific features of anchor nodes.*
- The *specific features of non-anchor nodes* of elementary trees. These features are not associated to the anchor of tree; rather they are connected to other types of nodes, which affect the meaning of the anchor node.

Automatic Generation of Decision Trees

All attributes mentioned in the previous section have been implemented using a data set with 2136 lexical items (including proper nouns) of 860 sentences shorter than 16 words. Using TAG formalism and after parsing these sentences, all attributes get their appropriate values, and the correct permutation of Persian words is generated. By aligning the related words of English and Persian sentences, the information required for building a decision tree by using an algorithm such as ID3 would be available [13].

There is a decision tree for every word that appears in the training set. These trees can be derived by using ID3, which chooses one attribute at any step to divide the training set. By using these attributes, the optimal decision trees that maximize the gained information, are generated.

Figure 3 shows the learned decision trees based on our training data set for two words: "can" (3.a) and "who" (3.b). The meaning of "*can*" is firstly determined by its part-of-speech tag. If it is a noun, the Persian word "*Kozeh*" is selected as its meaning, whereas if it is a verb, it may still either be an auxiliary or a main verb. This information is recognized by the elementary tree in which it is participated. The tree family B_2 is related to the auxiliary trees of phrases such as "*can hold*". Having this tree, the exact meaning of "*can*" can be determined.

The word "*who*" can be translated into three different Persian meanings: 1) it is used for making interrogative sentences. 2) it is used in relative clauses. 3) it plays the role of a common noun phrase.

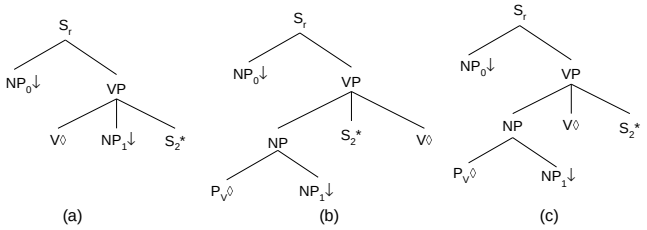

Figure 2. (a) The elementary tree of a declarative sentential complement with NP. (b) The first related Persian elementary tree (the main verb of the embedded sentence is in infinitive mode). (c) The second related Persian elementary tree (the main verb of the embedded sentence is in declarative mode.

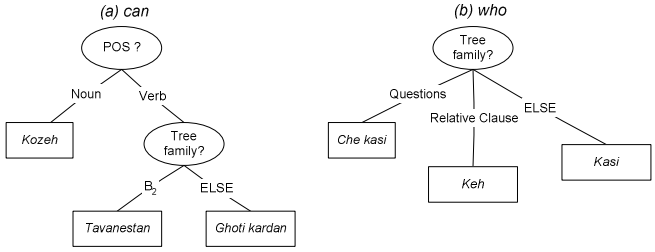

Figure 3. Decision trees for the lexical transfer of "can" and "who"

3.2.3 Feature Transfer Model

The third phase of transfer is the so-called feature transfer. Because of the divergences between Persian and English words, we need to transfer some features from English into Persian, which are later used to generate the proper form of individual words. These features can be divided into two types: *structural* and *morphological.*

Structural features are those that are related to a special structural form of Persian sentences. For example, the different *modes* of a verb in Persian are different from that of English. Thus, we need some methods for deriving these features from the English structure.

Morphological features are those that inflect word surface forms. There are some different morphological forms in Persian, which should be derived from English sentences by assuming the existence of their structural forms. The Persian morphological features are listed in [11]. In our experiments, all features could be derived from the English-Persian S-TAG. For example, there is a

[5] A family tree is a collection of semantically related elementary trees [17].

morpheme named *Enclitic Particle* in Persian that forces to attach a letter "*i*" to the end of a noun that is referred to by a relative clause. We extracted the value of this feature from all elementary trees related to the relative clauses; these trees are well-defined in the English-Persian S-TAG notation.

3.3 Morphological Rule Applier

The final phase of translation, which happened to be the simplest one, involves the application of Persian morphological rules. After performing feature transformation, some modifications are applied to individual words by using some morphological rules. The main property of this phase is the locality of its action: each rule applies only to a single word, and does not affect the long distant words. We have implemented all those morphological rules that have been explained in [12].

4 A COMPLETE EXAMPLE

In this section, we introduce a complete example of English-Persian translation. The sentence pair is:

(3) Our software solutions can create customized packages for your special needs

rah-ehal-ha-i vizheh ra	narm-afzar-i ma barai-e niaz-ha-i	mi-tavan-ad khas-e shoma	basteh-ha-i tolid kon-ad
solutions customized	*software for*	*our needs*	*can special your*	*packages create*

By applying a TAG-based parsing algorithm to the above English sentence, the derivation tree shown in figure 4 is obtained. Each node of this tree refers to an elementary tree with its lexical values associated to its anchor.

The list of elementary trees with their associated Persian elementary trees is shown in figure 5[6]. For every elementary tree, the prefix letter *F* refers to its associated Persian one. Note that, there are usages of *Noun* and *Adjective* adjunctions in the auxiliary trees β_1 and β_4. The Persian forms of these trees are generated by changing the order of their arguments. There is also a usage of auxiliary verb in auxiliary tree β_2. Persian supports these kinds of combinations with the same order and semantics. Preposition adjunction can be handled by using the auxiliary tree β_3. The word "*ra*" in the elementary tree $F\alpha_1$, refers to a preposition that is used after objects of some verbs.

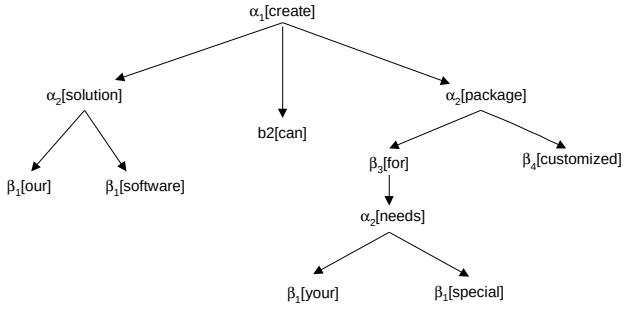

Figure 4. The derivation tree of sentence "Our software solution can create customized packages for your special needs"

Using the mentioned S-TAG, and by transforming the derivation trees shown in figure 5, we can derive the correct structure and order of Persian words. After these phases, all words are in their root form, which need to be inflected by the appropriate morphological rules. For example, there is a morphological rule called "*ezafe*" that attaches the suffix "-e" to the end of the first word of a compound NP.

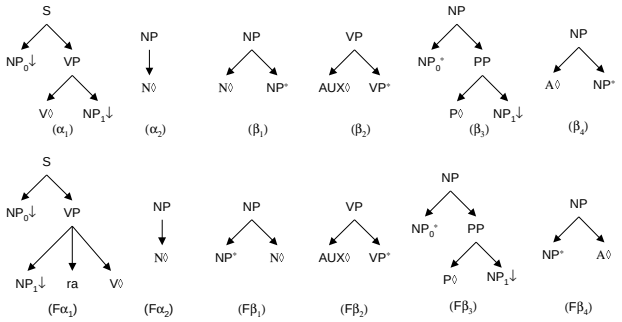

Figure 5. English and Persian elementary trees used for sentence (3)

5 CONCLUSION

We reported our implementation of an English-Persian translator by using the S-TAG formalism. Although we focused on translation from English into Persian, the ideas of this paper can also be applied to the reverse direction. However, the lake of a comprehensive computational grammar for Persian is the main obstacle in this regard.

In our experiments, we've used 860 sentences shorter than 16 words. For English TAG, we've extended the XTAG data set after selecting 150 elementary trees from the total number of 1227 elementary trees, which have been used for parsing the sample sentences [17].

By using the introduced transferred approach, we are now able to translate all sentences with different syntactical structures such as passive/active forms, and different tenses and persons.

Although S-TAG seems to be a suitable method for syntactic and semantic transformation, the little information that is used in the semantic interpretation of the context, makes it weak in encountering complex contexts. For example, there is no way to distinguish between the following sentences, by just using the S-TAG notion:

(4) "*John asked a man*"

John az	*yek*	*mard*	*porsid*
John	*a*	*man*	*asked*

(5) "*John asked a question*"

John yek soal		*ra*	*porsid*
John a	*question*		*asked*

The difference between these two translated sentences is related to special prepositions associated with direct/indirect objects. In Persian the preposition "*ra*" appears after a direct object, while some others preposition such as "*az*" is placed before an indirect object [4].

ACKNOWLEDGEMENTS

We would like to thank the anonymous reviewers of the paper for their invaluable suggestions and comments. This work is partially supported by the Iranian Telecommunication Research Center (ITRC).

[6] For simplicity, the Persian trees are shown in the reversed form (i.e., from left to right, similar to English).

REFERENCES

[1] A. Abeille and Y. Schabes, *Parsing idioms in tree adjoining grammars*, In Proceedings of the 4th Conference of the European Chapter of the Association for Computational Linguistics, Manchester, England, 1989.

[2] A. Abeille, Y. Schabes, and A. Joshi, *Using Lexicalized Tags for Machine Translation*, In Proceedings of the 13th International Conference on Computational Linguistics (COLING 90), pp. 1-6, Helsinki, Finland, 1990.

[3] 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, 2000.

[4] M. Bateni, *Tosif-e Sakhtari Zaban-e Farsi [Describing the Persian Structure]*, Tehran, Iran, Amir-Kabir Press, 1995

[5] E. de la Clergerie, M.A. Alonso Pardo, and D. Cabrero Souto, *A Tabular Interpretation of Bottom-up Automata for TAG*, In Proceedings of the 4th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+4), pp. 42-45, Philadelphia, PA, USA, 1998.

[6] E. de la Clergerie and M.A. Alonso Pardo, *A Tabular Interpretation of a Class of 2-Stack Automata*, In Proceedings of the 17th International Conference on Computational Linguistics (COLING 98) and the 36th Annual Meeting of the Association for Computational Linguistics (ACL 98), pp. 1333-1339, Montreal, Canada, 1998.

[7] B.J. Dorr, *Machine translation divergences: A formal description and proposed solution*, Computational Linguistics, **20**(4): 597-633, 1994.

[8] M. Dras and C. Han, *Korean-English MT and S-TAG*, In Proceedings of the 6th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+6), pp. 206-219, Venice, Italy, 2002.

[9] P. Khanlari, *Tarikh-e Zaban-e Farsi [History of Persian Language]*, Tehran, Iran, Simorgh Press, 1995.

[10] D. Mark and T. Bleam, *How problematic are clitics for S-TAG Translation?*, In Proceedings of 5th International Workshop on Tree Adjoining Grammars and Related Formalisms (TAG+5), pp. 241-244, Paris, France, 2000.

[11] K. Megerdoomian, *Unification-Based Persian Morphology*, In Proceedings of CICLing 2000, Alexander Gelbukh, Center of Investigation on Computation-IPN, Mexico, 2000.

[12] K. Megerdoomian, *Persian Computational Morphology: A unification-based approach*, NMSU, CLR, Memoranda in Computer and Cognitive Science Report, 2000.

[13] T. Mitchell, *Machine Learning, McGraw Hill*, 1997.

[14] S.M. Shieber and Y. Schabes, *Synchronous tree adjoining grammars*, In Proceedings of the 13th International Conference on Computational Linguistics (COLING 90), pp. 253-258, Helsinki, Finland, 1990.

[15] M. Stevenson, *Word Sense Disambiguation: The case for combining Knowledge Sources*, CSLI publication, Stanford, CA, 2003.

[16] G. Van Noord, *Head-corner parsing for TAG*, Computational Intelligence, **10**(4), pp. 525 – 534, 1994.

[17] XTAG Research Group, *A Lexicalized Tree Adjoining Grammar for English*, Technical Report IRCS 98-18, Institute for Research in Cognitive Science, University of Pennsylvania, pp. 5-10, 1998.