Example-based Machine Translation from English to Farsi with the help of WordNet Ontology

Hassan Shojae-Mend* and Mostafa Amiri

Department of Basic Sciences, Faculty of Medicine, Gonabad University of Medical Sciences, Gonabad, Islamic Republic of Iran; hassan.shojae@gmail.com, mu1amiry@yahoo.com

Abstract

Machine translation is a process of automatic translation from one language to another. In the present study, architecture was proposed for the system of example-based machine translation from English to Farsi with the help of WordNet ontology. This system uses a knowledge base containing a set of examples, a bilingual dictionary and WordNet. An effective factor in quality of translation is ambiguity in selecting one of the different meanings that a word can have while being translated from one language to another. In the proposed method of the present study, WordNet is utilized in the formation phase of example set and the adaptation phase so that ambiguity in translation can be reduced and selection of the correct meaning from among different meanings will be possible.

Keywords: Example-based Machine Translation, Ontology, Semantic Web

1. Introduction

Machine translation system refers to computer systems that automatically translate a text from one language to another. Although the main goal of a machine translation system is to produce a high quality translation without human interference, in practice this has been possible in limited situations and with low quality1. Machine translation systems can be divided into two general groups: rule- and knowledge-based machine translation systems. In rule-based machine translation systems, the attempt is to obtain the translation of an input sentence into the target language by utilizing the relationships available between linguistic rules of the source and target languages. Some problems of this method include impossibility of exact determination of linguistic rules and the relationships between them in different languages, ambiguities involved with translating words from one language to another and different meanings that words can have in translation.

Knowledge-based machine translation systems are classified into two general groups: example- and statistical-based machine translation systems. The pivotal idea behind them is to use previously-done translations in order to achieve the translation of a new sentence. In translating a sentence, individuals do not use linguistic analyses, but they first divide it into a set of phrases then translate them into the target language and combine the resulting phrases so that the final sentence can be obtained in the target language. Translation of the resulting phrases is also carried out using the analogy translation principle through previously-translated examples2. Example-based machine translation system is composed of three main components: parallel corpora, matching, and recombination.

WordNet is a lexical database for English including names, verbs, adjectives, and adverbs that are classified in the set of synonyms3. Words available in a set are close to one another in regard with their meaning. Moreover, the relationship among different meanings is expressed through a hierarchical structure.
2. Related Works

Semantic web is an extension to the current web in which meanings are also included in a way that computers and humans can have better cooperation. One of the applications of annotation methods like Resource Description Framework (RDF) that is proposed in semantic web makes it possible to extract different sentences and their equivalent translations from web pages and use them in example-based machine translation systems. Ontology can be utilized to maintain the structure of the input sentence, whereby more exact equivalents can be found through ontological alignment techniques from the parallel corpora. In their study, Barman et al used Assamese WordNet and obtained a better quality translation from English to Assamese based on the general concept available in the text and by replacing one word with synonyms.

WordNet has been used in different studies to resolve ambiguities of translations. It was utilized in translating from English to Hindi and English to Bengali in order to obtain the exact meaning of a word according to its concept.

3. The Proposed System

The proposed architecture for the example-based machine translation system with the help of WordNet ontology is presented in Figure 1. Ontology is used to store the examples. The information related to the parallel corpora in this ontology is expressed in RDF format. In addition to storing the grammatical information of the sentences in the parallel corpora, the meanings of different components of the sentences are extracted using WordNet and stored in the parallel corpora. Information related to the meanings helps resolve ambiguities of the sentences and different meanings that exist for a word from one language to another. Every component of the system will further be examined later in the following paragraphs.

3.1 Parallel Corpora

One of the components of an example-based machine translation system is the parallel corpora. Any example in the parallel corpora includes two sentences; one related to the English and the other to the Farsi text. Any of these sentences is broken into its components, and connection is made between the components of the English and Farsi sentences. The structure used in the database to store the examples and the method of breaking a sentence and doing the related calculations related to the relationship between different components of English and Farsi sentences are explained below.

Figure 1. The architecture of the example-based machine translation system with the help of ontology.

3.1.1 Information and Structure of the Examples

Information that is stored for any example if the database includes two groups: one related to grammatical structure of a sentence and the other related to its Farsi translation. Dependency analysis is utilized to obtain the grammatical structure of the sentence. In dependency analysis method, a sentence is broken down into its smaller components based on grammatical rules, and this process continues until unbreakable grammatical components will be achieved. This process forms a tree of the grammatical structure of the sentence, which is called dependency tree. A sample of dependency tree is presented in Figure 2.

To store the information related to the parallel corpora, ontology is used whose structure is presented in Figure 3. Any of the groups that exist in the dependency tree showed in Figure 2 is called a phrase. Any phrase can include no or some other phrases. Any phrase that is related to the English sentence is connected to a phrase related to its corresponding Farsi translation. This connection is made through the feature of “hasLink”. The
method of calculating the relationship between English and Farsi phrases is explained.

The role of the phrase in the sentence along with its part of speech is placed within the category. For every phrase, the part of speech is also determined in the dependency tree. Parts of speech can be SUBJ, OBJ, HEAD, NP, ADJ, V, N, and VP. A category can contain one or some of these parts of speech. To differentiate among English and Farsi phrases, “E” is added to English phrases and “F” to Farsi ones.

In the description section, Figure 3, English phrase or its Farsi translation is placed. For instance, any of “buys”, “on semantic web”, and “semantic” from the dependency tree of Figure 2 is considered as a description related to that phrase. For each phrase, whenever possible, the concept is extracted from WordNet, and reference to that concept is saved by storing the number related to it in “Sense”. The method of calculating the concept of a phrase is explained below.

3.1.2 The Relationship between English Phrase and its Farsi Equivalent

The initial English phrase that includes the whole sentence is connected to the initial Farsi sentence. For the rest of phrases, a bilingual English-Farsi dictionary is used. Meanings related to all of the words existing in the phrase are obtained from the dictionary. For all Farsi phrases, it is calculated that the meanings of how many English phrases are available in them. Afterwards, the extent of similarity between Farsi and English phrases is calculated using the following formula:

\[
\text{Similarity} = \frac{\# \text{ExistedFarsiWord}}{\# \text{EngishWord}}
\]

Where, \( \# \text{EngishWord} \) indicates the number of words in the English phrase and \( \# \text{ExistedFarsiWord} \) determines the number of words of the English phrase whose meaning exists in the Farsi phrase. An English phrase is connected to a Farsi one, which has a higher level of similarity. Typically, the similarity should be close to 1. The extent of similarity is stored in the system, and after formation of the parallel corpora, cases with low levels of similarity will be debugged manually.

3.1.3 The Method of Calculating the Concept of a Phrase using WordNet

WordNet is utilized to calculate the concept of a phrase. The target phrase is searched in WordNet, and all of its concepts will be extracted. Afterwards, the meanings related to the words obtained for each concept will be obtained from the bilingual dictionary. This collection of the Farsi words are compared with those available in the corresponding Farsi phrase, and the share will be calculated. Any concept whose number of words inside the calculated shared set is higher is considered as the concept of the phrase and the related number will be stored in the “Sense” related to the English phrase.

3.2 Matching

As an English sentence enters the system, it should be translated into Farsi using the database of the examples. First, the tree structure of the sentence will be created by dependency analysis method. As was explained in the previous section, the information of this tree is stored in the form of an ontology.
3.2.1 Finding the Match based on the Structure of the Sentence

Ontology arrangement is used to find the matches\(^1\). The level of similarity between the input phrase and the levels available in the database of the examples is calculated. The weighted sum method is utilized in order to calculate the level of similarity between the two phases. The weight related to the connection with “Category” and ‘Description” is set at 0.25. And the weight of all other phrases related to another phrase is set as 0.5. Therefore, a series of phrases will be obtained from the database of the examples, each of which has a similarity level of 0 to 1 to the input phrase.

Since the structure of the input sentence and the database of the examples are utilized in the method of finding similar phrases, it is possible that the similar phrases found are not conceptually appropriated to the input sentence. Therefore, in the next step, the concept of the available words should be taken into consideration in determining the phrase similar to the input sentence.

Example: Suppose, we want to translate the sentence, “I take a meal”. The following sentences are selected from the database as the similar sentences:

- I take a pen.
- I eat meal.

In the above example, selecting one of the two extracted sentences from the database requires utilization of the meanings of the words. If the first sentence is selected, the obtained meaning will be wrong. Therefore, the second sentence should be selected, which is only possible by using structural aspects of the sentence.

3.2.2 Limiting the Matches based on the Meaning of the Sentence

In order to limit the found matches, WordNet is utilized and the concept matching of the sentence is also taken into account. Phrases related to the found matches are arranged in a descending order, and then the first n phrase will be selected as the candidate phrase. Now, for each candidate phrase and for each word of that phrase, the code presented in Figure 4 is applied. This pseudocode calculates the similarity level between a word of an input phrase and its equivalent word in the found match of the database of the examples. The calculated digit will be between 0 and 1. Finally, the level of conceptual similarity for each phrase will be equal to the total conceptual similarity obtained for the words divided by the number of the words. Each phrase that has a higher level of conceptual similarity is selected as the final match.

Figure 4. The pseudocode for calculating the conceptual similarity between two phrases.

In the pseudocode of calculating the conceptual similarity, if the word is a proper noun or pronoun, the level of similarity between them is taken as 1. If the word is a noun, all concepts that the input word can have are calculated using WordNet. Next, the most depth of the tree for each of these concepts is calculated as far as the common parent node, then the obtained number is divided by the height of the tree as far as the target concept, and finally 1 is subtracted from the obtained number. Using these calculations, the closer the two concepts are to one another in the tree structure of WordNet, the closer to 1 their similarity will be, and the more distant they are, the closer to 0 the related digit will be. If the word is a verb, its equivalents will be extracted from WordNet, and if the input word exists in this set of synonyms, the conceptual similarity level will be 1 and otherwise 0.

In the previous example, the word “pen” in the first candidate phrase should be replaced with “meal”, and in the second phrase the word “eat” should be replaced with “take”. The depth of the node related to the concept ascribed to pen from WordNet to common parent node with the concept of “meal” is 7. The height of the tree up to that node is equal to 9. Therefore, the calculated level of conceptual similarity for these two words is 0.22. Since

\(\text{Function calculate-similarity (input-phrase, sub-phrase)}\)

\text{if sub-phrase-category is “noun” then}
\text{if sub-phrase-category is “proper-noun” then}
\text{if sub-phrase-category is “pronoun” then}
\text{if sub-phrase-category is “verb” then}
\text{if sub-phrase-category is “description” then}
\text{if sub-phrase-category is “category” then}
\text{calculate the similarity between input-phrase and sub-phrase}
\text{return sub-phrase-similarity}
\text{end function}
“eat” and “take” are both verbs and they are in the same synonym set in WordNet, the level of their similarity will be calculated as 1. Therefore, the level of conceptual similarity between the input sentence and the second candidate phrase is higher.

4. Evaluating the System

Since different correct translations of a single text can be provided by human, and words can be placed in many different places, the distance between the sentence translated by human and the one by machine cannot be measured clearly. A proposed method of evaluation is to count the number of lexical, syntactic, and semantic mistakes. In one of the automatic methods of evaluating machine translation, n-gram similarity factor and a series of candidate sentences along with different probabilities of translation are utilized. In this method, sentences along with appropriate translations should be selected as the test set.

In order to evaluate the proposed system, a test set including some sentences has been selected, which includes ambiguity in their translation. The method of resolving the ambiguity and achieving an appropriate translation through a proposed method have been investigated in the present study. The parallel corpora displayed in Figure 1 are considered for the system.

Table 1. Database of the examples

| Sentence                      |
|-------------------------------|
| Ali played the game           |
| John take a meal              |
| Monkey ate banana             |
| I read a book on computer     |
| You saw him on the bus        |
| Put it on the table           |

The sentence “Ali ate the game” is inserted to the system. Based on structural similarity, the following two candidate phrases are obtained from the database of the examples for this sentence.

- Ali played the game.
- You ate banana.

The first sentence is very similar to the input sentence while the second sentence is only similar in its verb. Using the mentioned method in section 3.2.2, the system calculates the conceptual similarity between the input sentence and these candidate phrases. The level of the obtained conceptual similarity is as follow:

- The conceptual similarity of the first candidate phrase with the input sentence: \((1+0+1)/3 = 0.75\)
- The conceptual similarity of the second candidate phrase with the input sentence: \((0.846+1+0.625)/3 = 0.823\)

Therefore, the second phrase is considered to be similar to the input sentence and the selected concept for “game” while calculating the level of similarity is taken as the concept of this word for translation, and the sentence is correctly translated into “Ali کهکشان کرد".

Ambiguity in translating prepositions is another issue in translation. For instance, different meanings that the preposition “on” can have are available in the examples of the above database. The input sentence, “I read a book on the car” is given to the system. Following sentences are extracted from the database of the examples:

- I read a book on computer.
- You saw him on the bus.
- Put it on the table.

From the first sentence, the phrase “I read” and “a book on computer” is similar with the phrases of “I read” and “a book on the car”.

From the second sentence, the phrase “on the bus” is similar to the phrase “on the car”.

From the third sentence, the phrase “on the table” is similar to the phrase “on the bus” from the input sentence.

The conceptual similarity of the first candidate phrase with the input sentence: \((1+1+1+0.45)/4=0.86\)

The conceptual similarity of the second candidate phrase with the input sentence: \((1+0+0+0.66)/4=0.41\)

The conceptual similarity of the third candidate phrase with the input sentence: \((0+0+0)/=0\)

Therefore, the first candidate phrase is selected as the phrase with the highest level of similarity with the input sentence. The preposition “on” in the input sentence is translated like the one in the first candidate sentence, meaning “about”.

5. Discussion and Conclusion

In the present study, a structure was proposed based on ontology in order to store syntactic and semantic information of a phrase in a knowledge base. By using WordNet and a bilingual dictionary, it is possible to discover and store the concepts of the words in sentences. Utilizing
WordNet along with an example-based machine translation system makes it possible to conceptually resolve ambiguities in translation in case of presence of different meanings for a word in the parallel corpora. In the proposed method to find similar sentences, in addition to structural match between the sentences, conceptual match is also utilized in order to resolve ambiguity. By calculating the conceptual match, many candidate translations that are conceptually different can be crossed out. Although the proposed method resolves many ambiguities of translation, there will still be cases of ambiguity in the translation. If the meaning of a sentence depends on the previous sentences in the text, this system fails to distinguish the concept related to the sentence in question. For instance, in the sentence, “What’s eating you?” meaning “What worries you?” the exact meaning cannot be determined without considering the sentences of the text. Moreover, the system will encounter problems in translating verbs whose subject should be replaced, resulting in a different translation for the verb. This problem can be resolved by adding some syntactic considerations to the system.

6. References

1. Xuan H, Li W, Tang G. An advanced review of hybrid machine translation (HMT). Procedia Engineering. 2012; 29:3017–22.
2. Arnold D, Balkan L, Humphreys RL, Meijer S, Sadler L. Machine translation: An introductory guide. London: NCC Blackwell; 1994.
3. Miller GA. WordNet: A lexical database for English. Communications of the ACM. 1995; 38(11):39–41.
4. Berners-Lee T, Hendler J, Lassila O. The semantic web. Scientific American. 2001; 284(5):28–37.
5. Language resources for the Semantic Web: Perspectives for machine translation. In: Vertan C, editor. Proceedings of the Second International Workshop on Language Resources for Translation Work, Research and Training; Association for Computational Linguistics; 2004.
6. Elita N, Birladeanu A. A first step in integrating an EBMT into the Semantic Web. 2005. Available from: http://www.mt-archive info/MTS-2005-Elita.pdf
7. Barman AK, Sarmah J, Sarma SK. Assamese WordNet based quality enhancement of bilingual machine translation system.
8. Resolving pattern ambiguity for english to hindi machine translation using WordNet. In: Chatterjee N, Goyal S, Naithani A, editors. Workshop on Modern Approaches in Translation Technologies; 2005.
9. Salm KMAS, Khan M, Nishino T. Example based English-Bengali machine translation using wordnet. 2009.
10. Kruijff G-J. A categorial-modal logical architecture of informativity: Citeseer; 2001.
11. Shvaiko P, Euzenat J. Ontology matching: state of the art and future challenges. IEEE Transactions on Knowledge and Data Engineering. 2013; 25(1):158–76.
12. Papineni K, Roukos S, Ward T, Zhu W-J. BLEU: a method for automatic evaluation of machine translation. Philadelphia, Pennsylvania: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Association for Computational Linguistics; 2002. p. 311–8.