Review

An adaptive approach of syntactic ambiguity resolution in Pashto

Muhammad Bilal*, Mohammad Abid Khan, Rahman Ali and Rashid Ahmed

Department of Computer Science, University of Peshawar, Peshawar, Khyber Pakhtoonkhwa, Pakistan.

Accepted 22 December, 2010

A natural language contains a variety of ambiguities. From a computational point of view, it is very difficult to cope with these ambiguities. Syntactic ambiguity arises from the relationship between the words and clauses of a sentence, not from the range of meanings of single words. Here, a knowledge-based adaptive approach has been proposed to resolve the syntactic ambiguities in Pashto text. The resolution involves user interaction with the system to select the appropriate meaning of the ambiguous phrase from a set of possible meanings stored in the knowledge base of the system. The user can add a new meaning to the knowledge base, for a particular phrase, if it is not there. The system has been tested on a raw Pashto corpus and resulted in accuracy of 90%.

Key words: Ambiguity, Pashto, syntactic ambiguity, syntactic ambiguity resolution, algorithm.

INTRODUCTION

Ambiguity is an inherent characteristic of natural languages (Bilal et al., 2009; Attia, 2008). It is a persistent phenomenon in almost all the areas of natural language (NL), that is, semantics, phonology, morphology, syntax, and it has been massively explored in language processing at least since Bever (1970) and Erdocia et al. (2009). When most of the words in any NL text are seen in isolation, the intended meaning cannot be determined. It can only be determined by applying some contextual, probabilistic or real world knowledge clues (Attia, 2008).

Ambiguity resolution has long been the focus in natural language processing (NLP) (Bilal et al., 2009; Su et al., 1990). For a native speaker of a particular NL, to resolve the ambiguity is a very tedious task. The computational analysis of human language is even more complicated, as there arise a lot of other ambiguities, besides the real ambiguities, due to the interaction of rules made for resolving these ambiguities (Attia, 2008). It remains one of the main problems that arise in NLP (Blache, 1996).

For the resolution of ambiguities two types of approaches are used by the researchers mostly: two-stage theories and constraint-based theories (Gompel et al., 2000). According to two-stage theories presented by Frazier (1979) and Rayner et al. (1983), when an initial adopted meaning of an ambiguous input is inappropriate then reanalysis should occur. Whereas, in constraint-based theories presented by MacDonald (1994) and McRae et al. (1998), the authors claim that processing difficulty is due to a competition between two or more syntactic analyses that are about equally activated.

Van Gompel et al. (2000) claimed that both the above theories of sentence processing have been ruled out after their reviewing of several experiments investigating this issue. They proposed another solution for performing the sentence processing: The unrestricted race model (Gompel et al., 2000). This model combines properties of both constraint-based and two-stage models, but is different from both. In this model, the alternative structures of a syntactic ambiguity are engaged in a race, with the structure that is constructed fastest being adopted (Gompel et al., 2000).

In different languages, much work has been done for the resolution of ambiguities: these include English, Chinese, Basque, Arabic, German, Japanese (Sturt et al., 2002; Li, 2003; Erdocia et al., 2009; Attia, 2008; Kevin, 2001; Skut et al., 1998; Fodor et al., 2003), and a lot more. Several approaches have been applied for the resolution process of ambiguities. These include probabilistic, knowledge-based and rule-based approaches (Bilal et al., 2009). This paper is about the syntactic ambiguity resolution in Pashto language text. Pashto is one of the Indo-Iranian Arabic script languages (Rahman, 1995). It is

*Corresponding author. E-mail: mbilal82@gmail.com.
spoken in Afghanistan, Pakistan and some areas of India. It is one of the richest languages of the world and like all other natural languages it has the problems of ambiguity, anaphora and ellipses (Bilal et al., 2009; Ali, 2008a; Ali, 2008b; Ali, 2008). Work has been done on the morphological aspects of Pashto language as well, which includes the morphological structures of Pashto nouns and verbs (Zuhra and Khan, 2008; Zuhra and Khan, 2007; Khan and Zuhra, 2007). This work is the first effort as before this work was done in Pashto for the problem of ambiguity resolution. The approach used in this work is the knowledge-based approach. A raw Pashto corpus with partially tagged syntactically ambiguous examples is scanned. If an ambiguous phrase is found during the scanning, its meaning is first searched in the resolution table (RT). The RT contains all those solutions which are being opted by the users in the past. If the actual meaning of the ambiguous phrase is not present in the RT, or the meaning in the RT is not the correct one, then all the possible meanings of that particular phrase are prompted to the user from the possible meanings table (PMT). The user then selects the actual meaning from the PMT. If there is no solution found in the PMT for a particular syntactically ambiguous phrase, then the user can add his possible meaning of that particular phrase in the RT. The new meaning added to the RT is also added in the PMT with the identification mark of the particular phrase for which the meaning is added. So, the knowledge-base could be built, in terms of possible meanings, with time. This is an adaptive approach in which the system adapts the newly added meaning for the future use.

AMBIGUITIES IN PASHTO

Natural languages are inherently ambiguous as discussed earlier. The types of ambiguities that are found so far in Pashto language are three: Lexical ambiguity, Syntactic ambiguity and Pragmatic ambiguity (Bilal et al., 2009). Lexical ambiguity occurs when a word can have multiple meanings either individually or if that word comes in a particular phrase. A phrase or sentence will be having a syntactic ambiguity when there are multiple related grammatical structures made for it. Pragmatic ambiguity occurs when the speaker and the listener do not agree on the same principals of communication in a particular language (Bilal et al., 2009; Jurafsky and Martin, 2000). In this paper, the emphasis is on the syntactic ambiguities in Pashto text and their resolution.

SYNTAX AMBIGUITIES

A sentence or phrase is syntactically ambiguous when a sequence of words is compatible with more than one grammatical structure (Long et al., 2008). A native speaker of a particular NL can easily cope with such type of ambiguities unless the ambiguous text is written or spoken without context. From computational point of view, the sentence processing is done with the help of some grammatical rules made for the identifications of words, phrases and sentences. This identification is independent of any context. So, after the processing, the system will declare all those phrases ambiguous, for which there exist more than one valid grammatical structure.

Syntactic ambiguities in Pashto language are of three different types. These include Attachment ambiguity, Coordination ambiguity and Idiomatic-Verb-Phrase ambiguity as identified by Bilal et al. (2009). A proposed rule-based approach has been proposed for the identification of all these syntactic ambiguities (Bilal et al., 2009). In the identification process, a partially tagged set is passed through a manual parsing mechanism as there is no parser developed so far for Pashto language (Bilal et al., 2009). The phrases for which more than one parse tree is generated is passed through the rule-based system which actually classify the phrase as one of the above syntactic ambiguity.

SYNTACTIC AMBIGUITY RESOLUTION

Syntactic ambiguity resolution is a central issue in NLP. A sentence is syntactically ambiguous if it can be represented by more than one grammatical structure (Attia, 2008). Syntactic ambiguity results when the text is encountered word by word during the sentence processing (Green et al., 2006).

The resolution process, defined in this paper, is for the syntactic ambiguities that arise in the Pashto text. The method proposed is a knowledge-based one. We have developed a knowledge base with all the possible meanings of the ambiguous phrases that appeared in the raw corpus.

The syntactic ambiguities in the corpus are identified with the system proposed by Bilal et al. (2009). All the possible solutions of an ambiguous phrase are saved in the knowledge base with the identification mark of the ambiguous phrase. When the system finds the phrase as ambiguous, then the meaning of that phrase is prompted to the user on the screen from the RT. RT contains the meanings of the syntactically ambiguous phrases which are being selected by different users of the system in the past. If the meaning shown to the user from RT is not the one that the user intended then the user has the option to check the meaning in the PMT. PMT contains all the possible meanings of each syntactically ambiguous phrase. The system prompts all the possible meanings to the user on the screen. Now, the choice is of the user to select the intended meaning of that ambiguous phrase from PMT.
The above method is explained with the help of some examples from the raw Pashto corpus. Example 1 is a syntactically ambiguous phrase from real Pashto text.

Example 1

Example 1 is syntactically ambiguous with three different meanings although the grammar is valid in all the three cases. The system shows the meaning from the RT, which will be one from the following meanings.

**Meaning 1 of Example 1**

<table>
<thead>
<tr>
<th>Me</th>
<th>this much</th>
<th>[coquet]</th>
<th>[do]</th>
<th>[and]</th>
<th>[you]</th>
<th>[even once]</th>
<th>[do]</th>
<th>[no]</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ze</td>
<td>[dwmre]</td>
<td>[nxri]</td>
<td>[kwm]</td>
<td>[aw]</td>
<td>[tasw]</td>
<td>[aw]</td>
<td>[kwey]</td>
<td>[nə]</td>
</tr>
</tbody>
</table>
| "I use to coquet all the time and you do not do it even once".

**Meaning 2 of Example 1**

| Me | [when] | [like] | [coquet] | [do] | [you] | [that] | [like] | [coquet] | [no] | [do] | [za] | [sng] | [nxri] | [kwm] | [aw] | [tasw] | [aw] | [kwey] | [nə] |
| "You do not coquet as I do".

**Meaning 3 of Example 1**

| Like | [that] | [me] | [coquet] | [no] | [do] | [you] | [too] | [no] | [do] | [sng] | [za] | [nxri] | [nə] | [kwm] | [tasw] | [hem] | [nə] | [kwey] |
| "You do not coquet as I do not".

Basically all these three meanings of the particular phrase, that is, Example 1 are placed in the PMT. If the meaning shown to the user from the RT is not according to the user’s will then all these three meanings are shown to the user from the PMT.

When the system finds a syntactically ambiguous phrase, it checks the knowledge base with the identification mark of this particular phrase. All the meanings which match the identification mark are then prompted to the user for the selection of the actual meaning in the respective context of the ambiguous phrase. In case of Example 1, the identification mark matches with three different meanings and are shown to the user.

There could also be some odd situations in which one is that: the RT or PMT do not have the intended meaning which the user wants and the second situation is that: the discourse is added with new syntactically ambiguous phrases for which there are no possible solutions in PMT or RT. For handling these types of situations, the user will be prompted to add his/her own meanings to that particular syntactically ambiguous phrase, which will be ultimately stored to the knowledge bases, that is, PMT and RT.

Let us consider another syntactically ambiguous phrase in Example 2.

Example 2

Example 2 has also two different meanings:

**Meaning 1 of Example 2**

| This | [when] | [happen] | [do] | [that] | [me] | [with] | [Gulmeena] | [with] | [no] | [this] | [when] | [happen] | [do] | [lwbe] | [aw] | [bia] | [do] | [gulmne] | [sra] | [na] | [da] | [kla] | [kedi] | [šw]
| "It was not possible for me, not to play with Gulmeena".

The syntactically ambiguous example 2 also has two different meanings:

**Meaning 2 of Example 2**

| This | [when] | [happen] | [do] | [that] | [me] | [with] | [Gulmeena] | [with] | [play] | [no] | [do] | [da] | [črta] | [kedl] | [šw] | [ča] | [ma] | [da] | [gulmne] | [sra] | [lwbe] | [nə] | [kwl] | [šw]
| "It would never have happened that I will not play with Gulmeena".

**WORKING OF THE SYSTEM**

The front end of the proposed system is developed in Microsoft C#.Net and the knowledge base is developed in Microsoft SQL Server. The interface of the system is shown in the system comprises of three tables named as AmbText, RT and PMT, shown in Figures 2, 3 and 4 of Appendix respectively. The syntactic ambiguities are identified with the help of the system proposed by Bilal et al. (2009), and are then tagged manually, to be identified.

The proposed system basically resolves the syntactically ambiguous phrases that happen to come in the partially tagged Pashto corpus. The knowledge base of the system comprises of three tables named as AmbText, RT and PMT, shown in Figures 2, 3 and 4 of Appendix respectively. The syntactic ambiguities are identified with the help of the system proposed by Bilal et al. (2009), and are then tagged manually, to be identified.
Table 1. The list of all abbreviations used in the algorithm.

<table>
<thead>
<tr>
<th>Abbreviation</th>
<th>Description</th>
</tr>
</thead>
<tbody>
<tr>
<td>Amb-no</td>
<td>Ambiguity number</td>
</tr>
<tr>
<td>EOF</td>
<td>End of file</td>
</tr>
<tr>
<td>MT</td>
<td>Machine translation</td>
</tr>
<tr>
<td>NL</td>
<td>Natural language</td>
</tr>
<tr>
<td>NLP</td>
<td>Natural language processing</td>
</tr>
<tr>
<td>PMT</td>
<td>Possible meanings table</td>
</tr>
<tr>
<td>RT</td>
<td>Resolution table</td>
</tr>
<tr>
<td>Sol.Id</td>
<td>Solution ID</td>
</tr>
<tr>
<td>SynAmb</td>
<td>Syntactically ambiguous</td>
</tr>
<tr>
<td>SynAmbText</td>
<td>Syntactically ambiguous text</td>
</tr>
</tbody>
</table>

by the system. When the system scans the corpus and finds the ambiguous phrases, it prompts all these phrases on the screen as Ambiguous Terms, shown in Figure 5 of Appendix, and are saved at the backend in a table named as AmbText, shown in Figure 1 of Appendix. At the front end, shown in Figure 5 of Appendix, the ambiguous phrases are highlighted in two colors: green and blue. The phrases which are highlighted in green indicate that this particular phrase has been resolved earlier. The phrases, highlighted in blue, are not resolved yet. The difference is that when the user wants to see the possible meanings of a phrase, highlighted in green, then the possible meaning list will show only one meaning which was being selected earlier. Alternatively, for the phrase highlighted in blue color, all the possible meanings will be shown. Both of these cases are shown in Figures 6 and 7 of Appendix respectively.

In Figure 6 of Appendix, if the meaning shown to the user is not acceptable by the user in his own context then he can check all the possible meanings by clicking the “Show All Possible Solutions button”, and ultimately selecting his intended meaning for that particular phrase. If in future another user wants to resolve the ambiguity for the same phrase, then the system will prompt the latest resolved meaning for that particular phrase. All these resolved ambiguities are stored in the knowledge base at the back end in the RT table, shown in Figure 3 of Appendix.

Figure 7 of Appendix shows all the possible meanings, which in this case are two, of a phrase in the possible solutions list. The user will select his intended meaning from the possible solutions list and that meaning will then be added to the knowledge base in the RT table.

If there comes a case in which the intended meaning for a particular phrase is not present in the PMT, then the user can add his own meaning to the knowledge base which will be ultimately added to the PMT and the RT table. This scenario is shown in the Figure 8 of Appendix.

For a newly identified syntactic ambiguity for which there is not any possible meaning in the knowledge base, the same procedure will be followed for adding the meanings as shown in Figure 8 of Appendix.

The final result is shown in Figure 9 of the Appendix, in which the ambiguous phrase is shown with the meaning that is selected by the user as the valid resolution. This resolved meaning can be used for any natural language processing system.

ALGORITHM AND PROGRAM ARCHITECTURE

An algorithm has been proposed for the resolution of syntactic ambiguities in real Pashto text. The algorithms takes partially tagged syntactically ambiguous phrases from the corpus and add these phrases to the SynAmb table of the knowledge base. The other two tables, RT and PMT, contain the recently resolved meaning of a particular phrase and all possible meanings of that phrase respectively. A module for adding meanings for newly identified syntactic ambiguity is also present in the algorithm. The list of all abbreviations used in the algorithm and in the paper is given in Table 1.

[Read Raw Pashto Corpus]
While(~EOF)
Read Text
  if(Text Seemed = SynAmbTxt) Then
    move (SynAmb-Table)
    if(SynAmbTxt = SynAmb-table.SynAmbTxt) move (RT-Table)
    Prompt the Meaning to the User
    Else
      Move (PMT-Table)
      Label
      Select PMT-table.Meaning where
      SynAmb-table.Amb-no = PMT table.Amb-no
      if (PMT-Table.Amb-no = SynAmb-Table.Amb.no)
      Prompt to the User and Select the appropriate one
      Add (RT-Table)
      Else
        Manually Add All Meanings to (PMT-Table)
      Goto Label
    End if
  End if
End While

Working of the proposed algorithm is pictorially described in the form of a flowchart in Figure 10 of Appendix. Here, different modules of the algorithm are shown in data flow sequence. The output of the algorithm...
contains a list of syntactically ambiguous phrases with its resolved meanings as selected by the user.

TESTING AND RESULTS

The proposed algorithm was tested on a Pashto raw corpus containing different types of syntactic ambiguities. Out of the corpus, 150 phrases were extracted for the testing of the system. These phrases were manually tagged in the corpus and identified and stored by the system in SynAmb table which makes the knowledge base of the system. The algorithm correctly resolved 135 phrases that were syntactically ambiguous. Thus, the accuracy rate of the algorithm is 90%. The remaining 10% error rate is mostly due to the un-identification of the ambiguous phrases due to the un-availability of annotated corpus, phrase extractor, Pashto parser and lack of real world and domain knowledge of the text.

CONCLUSION

This work is about the resolution of syntactic ambiguities in Pashto language. For the automated resolutions of these ambiguities, an algorithm has been proposed and implemented which correctly identifies these ambiguities with a success rate of 90%. The approach used is an adaptive knowledge base approach in which the system improves itself by the frequent interaction with the user of the system. The system contains a large knowledge base of syntactically ambiguous phrases of the raw Pashto corpus and all possible meanings of each and every phrase.

REFERENCES

APPENDIX

Figure 1. Interface of the system.

Figure 2. Table showing ambiguous text.

Figure 3. Table showing the resolution number of ambiguous phrases.
Figure 4. Table showing all possible meanings.

Figure 5. Ambiguous phrases at front end of the system.

Figure 6. Ambiguous phrases with a meaning from RT.
Figure 7. Ambiguous phrases with meanings from PMT.

Figure 8. Adding a meaning to the knowledge base.

Figure 9. Results of the system.
Figure 10. Flowchart of the system.