Probabilistic Context Free Grammar for Urdu

Neelam Mukhtar1 Mohammad Abid Khan2 Fatima Tuz Zuhra3
Department of Computer Science, University of Peshawar, Khyber Pukhtoonkhawa, Pakistan
1 sameen_gul@yahoo.com
2 abid_khan1961@yahoo.com
3 fateeshah@yahoo.com

Abstract

A Probabilistic Context Free Grammar (PCFG) for Urdu is developed from an already developed Context Free Grammar (CFG) for sentences and phrases. Probabilities are assigned to the rules and two new terms i.e. special weights and special probability are introduced and assigned to few rules. Weights are assigned to rules after performing calculations based on observing the occurrences (frequency) of the second term on the right hand side of a particular rule. Furthermore, if a rule has zero frequency at present but in future it is expected to be used, then instead of assigning zero probability a small value (0.0001 in our case) is assigned to it. These rules (with special weights and special probability) are added like other rules to the Urdu PCFG. A PCFG for Urdu is thus obtained.

1. Introduction

Statistical parsers are gaining popularity day by day due to their accuracy and efficiency. A number of different statistical parsers are already developed (Collins, 1999; Charniak, 2000; Petrov et al., 2006). The main idea behind any statistical parser is to assign probabilities to the grammatical rules. "However, in practice, the probability of a parse tree being the correct parse of a sentence depends not just on the rules which are applied, but also on the words which appear at the leaves of the tree" (Lakeland and Knott, 2004).

Two main reasons are making Urdu a challenging language: first, its Perso-Arabic script and second, its morphological system that is having inherent grammatical forms and vocabulary of different languages such as Arabic, Persian and the native languages of South Asia (Humayoun, Hammarström and Ranta, 2007).

In Urdu, research is done from different point of views such as creating an Urdu corpus (Samin, Nisar and Sehrai, 2006; Becker and Riaz, 2002) and tagging the Urdu corpus (Anwar, Wang, Luli and Wang, 2007). Researchers have proposed different tagsets for Urdu whose number of tags is ranging from 10 (Schmidt, 1999) to 350 (Hardie, 2003). Now, one of the demanding areas for research in Urdu (from computational linguistics point of view) is parsing a corpus. Up to our knowledge, considerable amount of work has not yet been done in this area. An efficient and accurate parser is needed to parse Urdu corpus.

As probabilistic parsers have the property of efficiency and accuracy, so a probabilistic parser is needed to parse Urdu sentences. Before developing such a parser, a Context Free Grammar (CFG) is a pre-requisite. Furthermore, if the aim is to develop a probabilistic parser, first a probabilistic context free grammar (PCFG) is required. "A probabilistic context-free grammar (PCFG) is a CFG with probabilities assigned to grammar rules, which can better accommodate the ambiguity and the need for robustness in real-world applications" (Tu and Honavar, 2008). Tree-Bank based probabilistic grammar for Urdu is developed (Abbas, Karamat and Niazi, 2009). Now, PCFG is developed for Urdu by taking Urdu tagged sentences from different sources mentioned below. The development of PCFG is discussed in detail and the development steps are divided into different sections. Work in different sections of the research paper is organized as follows:

In section 2, text tagging is discussed. In section 3, the steps taken for developing a context free grammar are presented. In section 4, the process of the development of PCFG (including the potential

problems) is chalked out. In section 5, conclusion and future work is given.

2. Text tagging

Although some tagged text is made available by Center for Research in Urdu Language Processing (CRULP) under Urdu-Nepali-English Parallel Corpus project, but most of the sentences in this text are complex from parsing point of view. Therefore, apart from taking some complex sentences from the tagged corpus by CRULP (www.crulp.org), some more data was also collected. Specifically, the focus was on Urdu aqwaal-e-zareen, mazameen and mini-kahanian written by famous authors such as Saadat Hassan Minto, Ibne Inshah and Pitras Bukhari. Text was POS tagged by utilizing the annotator provided by CRULP. The tagset with 46 tags was used for text tagging. This tagset is developed recently by CRULP as part of a project for developing Urdu-Nepali-English parallel corpus (http://www.crulp.org/software/ling_resources/UrduNepaliEnglishParallelCorpus.htm). It follows the Penn Treebank guidelines. It has 46 tags e.g. grades of pronouns (PR, PRP$, PRRF, PRRFP$, and PRRL), demonstratives (DM and DMRL), several tags for verbs (VB, VBI, VBL, VBLI, and VBT), tags for auxiliaries showing aspect (AUXA) and tense (AUXT), NN tag for both singular and plural nouns, several other grades of common nouns (NNC, NNCR), two shades of Proper Nouns (NNP, NNPC) and a tag WALA which is used for every occurrence (and inflection) of word “wala”(Muaz, Ali and Hussain, 2009).

3. Context Free Grammar for Urdu

After having tagged data, Context Free Grammar (CFG) for Urdu was developed. Presently, a Pashto chart parser was utilized according to the requirements of Urdu language (Khan and Zuhra, 2010). This Pashto Chart parser can be used for any language. It was easily available as it is indigenously developed. So it was utilized in this work. Sentences were selected from the sources mentioned above and subsequently tagged. Moreover, tagged sentences by the CRULP (www.crulp.org) were also taken. Sentences were taken one by one and rules were written for each sentence using the abbreviations such as S (sentence), NP (noun phrase), PP (prepositional phrase) and VP (verb phrase). The rules were formed for identifying the boundaries of phrases only. The sentences were parsed by the chart parser and the result along with the rules was stored in a separate file.

After testing 200 sentences, majority of the rules started repeating themselves, but for maximum accuracy additional 100 sentences were taken and they contributed further few rules. The process was stopped when effort in the input text and parsing time reached up to un-feasible extent of either not delivering further rules or with very high time overhead. As examples from real text were used for the construction of rules, basic word order and free word order both were experienced. The developed grammar contains rules for both word orders that are mentioned above. A sample of the tested sentences, their corresponding rules and the output of the Chart parser are shown below:

Sentence	Rules	Results
<s> <w POS="NNP">ابن</w> <w POS="NNPC">انشاء</w> </s>	G->S S->NP NP->NNP NNPC	NP->NNP NNPC@-1,3
<s> <w POS="PRP1">ہمارا</w> <w POS="NN">ملک</w> </s>	G->S S->NP NP->PRP1 NN	NP->PRP1 NN@-1,3
<s> <w POS="NN">ہوش</w> <w POS="NN">رزق</w> <w POS="CM">کو</w> <w POS="VB">کھا</w> <w POS="VB">جاتا</w> <w POS="VBT">ہے</w> </s>	G->S S->NP VP NP->NN NP NP->NN VP->CM VP VP->VB VP VP->VB VBT	NP->NN NP@-1,2 S->NP VP@-1,7 S->NP VP@-1,7 NP->NN NP@-2,3 S->NP VP@-2,7 S->NP VP@-2,7 VP->CM VP@-3,7 VP->VB VP@-4,7 VP->VB VBT@-4,7 VP->VB VP@-5,7 VP->VB VBT@-5,7

Problems faced in Context Free Grammar and proposed Solutions

A $ sign in the tags (e.g. PRRFP$) was not accepted by the parser so it was replaced by ‘1’ whenever it appeared. Out of 300 sentences, 270 sentences were parsed successfully while 30 sentences failed to parse. So, the success rate of parsing is 90%. The rules for these 270 sentences were collected and normalized to form a CFG. The failed 30 sentences were analyzed carefully and the difficulties were mostly of the following nature:

1. In case of long and complex sentences when there were a number of rules and some rules were of the type
    1. NP->NN NP
    2. NP->NN VP
    3. NP->NN PP

   the parser failed to parse the sentence. The reason is that the rules have NP->NN in common but at the end, each rule is different (resulting in ambiguity) so it is difficult for the parser to decide as to which of the three rules to take next.

2. Parser usually failed due to left recursion in case of the rules such as S->VP VP.

The failed 30 sentences and their rules were also stored in a file for future analysis (to find out that whether any other parser is able to handle those sentences).

4. Probabilistic Context Free Grammar for Urdu

A PCFG is a CFG where each production is assigned probability. This probability is assigned to each rule by the simple formula:

P= Number of occurrences of a rule / Total Number of occurrences.

For calculating probabilities of the rules, the successive rules (that were used in the construction of the sentences) were arranged such that the rules for each of NP, VP, PP and S the rules were kept in separate files. The frequencies for NP, VP, PP and S were 2136, 914, 179 and 2091 respectively. These files were subsequently used to get the frequency of individual rule and ultimately to get the probabilities by the formula discussed above e.g. if the frequency of the rule S->VP NP is 23 and the total frequency of S rules is 2091 as already mentioned, then the probability of this particular rule is 23/2091=0.0109.

A section of the table showing the calculations of probabilities of NP is given below:

S.NO	RULE	FREQUENCY	PROBABILITY
1.	NP->NN NP	496	0.2322
2.	NP->NNP NP	147	0.0688
3.	NP->NNPC NP	35	0.0164

Problems faced in PCFG and proposed Solutions

Some problems were noticed in the development of PCFG but solutions were provided for these problems. These solutions are in the form of assigning special weights and special probability.

a. The rule such as VP->VB VBT (where VB and VBT are tags used for “Verb” and “Verb to be” respectively) was replaced by another rule VP->VB VP because of the existence of the rule VP->VBT. This rule (i.e. VP->VBT) was not directly used in the 300 sentences that were selected from the sources mentioned above. According to probability calculations, its probability will be zero. However, this rule may be used directly in other sentences, when the software (e.g. Urdu Probabilistic Parser) will be applied to huge natural text. A rule with zero probability contributes nothing to further calculations performed (though being used in the process of probabilistic parsing). These calculations are:

1. Finding the total probability of all the successive rules.
2. Selecting the total probability with the maximum value, as the most suitable parse of the sentence by Urdu Probabilistic Parser.

All the calculations are dependent on the probabilities of different rules. To avoid the problem of occurrence of a rule with zero probability, special weights were assigned to all such rules, but with few limitations. It should be noted that special weights can be assigned to the rules subject to the following conditions:

1. These particular rules are having a single term on the right hand side of the rule (e.g. in the rule NP->JJ, JJ is a single term on right hand side).
2. The single term on the right hand side of such rule should not be any phrase (e.g. NP, VP and PP) or sentence (S).
3. The only term on the right hand side of such rule should occur as a second term in other remaining rules at least once.

The special weights were calculated separately by the following procedure:

For NP, first the occurrences of JJ (adjective), Q (quantifier), RB (adverb), NNPC (proper noun continue) and NNCR (combined noun continue) in the second position on the right hand side of a rule were counted separately. Part of speech in the second position in the rule e.g. NP->NN RB is RB, so RB is in the second position. The number of times RB is appearing in this position (in all successive rules) was counted separately. After counting their occurrences (frequency), they were assigned special weights. Weights were calculated by two methods:

1. Dividing the occurrence of a particular tag (for example RB) by the total number of occurrences of all such tags (JJ, Q, RB, NNPC and NNCR) in the second position.
2. Dividing the occurrence of a particular tag by the total number of rules for NP which is 2136.

The same procedure was adopted for VP and PP.

A section of the table showing the calculations of special weights is given below:

Special Rules For VP	No Of Occurrences In The Second Position Of other Rules	Weights Based On		
		Percentage	Special Weights	
			Col II/245	Col II/914
VP->VB	17	6.94	.0694	0.0186
VP->VBL	12	4.89	.0489	0.0131
VP->VBT	142	57.96	.5796	0.1554

The weights calculated by the second method (i.e. last column in the above table) are providing smaller values than the first method. These weights (that are calculated by the second method) are used in PCFG as the aim is to use the smallest possible value (for accuracy) instead of assigning zero probability. These weighted rules can be used normally in parsing exactly like all other rules in probabilistic context free grammar that are assigned probabilities.

b. There were few rules (e.g. S ->VP, S->VP PP) that were expected to be used in future but at present they were not appearing in the sentences that were tested through the chart parser. Their frequency was zero thus resulting in a zero probability for these rules. The procedure used for assigning special weights was not possible to apply because of the following reasons:

In 9 rules (out of 12 such rules in total), there were two terms on the right hand side of the rule (e.g. NP->NP NP), which is violation of condition 1 mentioned in section a. In one rule (i.e. S->VP), the term on the right hand side of the rule was a VP, resulting in violation of condition 2. The right hand side of the remaining two rules (i.e. NP->PR and VP->VBI), were not occurring as a second term in any other rule which is violation of condition 3. It was not possible to assign special weights to these twelve rules.

To avoid this problem (i.e. including these 12 rules in PCFG with non zero probability), all the twelve rules mentioned above were assigned small probability i.e. 0.0001. As we are taking probability values up to four decimal places (in Urdu PCFG), so this is the smallest possible value in this case. Any new rule required in future can be added to the PCFG with this special probability (with a very small effect on accuracy).

A section of the table showing special probabilities, assigned to rules, is given below:

S.No	RULE	PROBABILITY
1	S->VP	0.0001
2	S->VP PP	0.0001
3	S->S S	0.0001

The benefits of assigning special weights and special probability to few rules are:

1. Addition of rules (that may be required in future) with non zero probability to the Urdu PCFG.
2. Special weights are helpful in reducing the number of rules in Urdu PCFG. We have a rule VP-> VBL with special weight. We already have the rules VP->JJ VP, VP->RB VP and VP->ITRP VP in Urdu PCFG. Now if we have 3 new rules i.e. VP->JJ VBL, VP->RB VBL and VP->ITRP VBL then instead of adding these 3 new rules to Urdu PCFG, they are merged in already existing rules. The rule VP->JJ VBL is merged in VP->JJ VP; VP->RB VBL is merged in VP->RB VP and VP->ITRP VBL is merged in VP ->ITRP VP by using the rule VP->VBL.

A section of the table showing Urdu PCFG is given:

S.No	Rules	Probabilities
1.	S->NP VP	0.9637
2.	S->PP VP	0.0167
3.	S->PP NP	0.0019
4.	S->VP NP	0.0109

Urdu PCFG is developed with 126 rules, including 11 rules with special weights and 12 rules with special probabilities assigned.

5. Conclusion and future work

Chart parser is used for accepting POS tagged text and displaying the rules that are used in the process of parsing. First a CFG and subsequently a PCFG for Urdu is developed. This PCFG can be used by the probabilistic parser for Urdu (that is to be developed)

that accepts POS tagged text as input and generates the structure of that text.

References

A.Hardie, The computational analysis of morph syntactic categories in Urdu, Ph.D thesis, Lancaster University. 2003a.

A.Muaz, A.Ali, S. Hussain, "Analysis and Development of Urdu POS Tagged Corpus", Proceedings of the *7th Workshop on Asian Language Resources*, *ACL-IJCNLP* 2009, pages 24–31.

C.Lakeland, A. Knott, "Implementing a Lexicalised Statistical Parser", Proceedings of the Australasian Language Technology Workshop, Macquarie University, Sydney, 2004.

D.Becker and K Riaz. "A Study in Urdu Corpus Construction", Proceedings of the 3rd Workshop on *Asian language resources and international standardization*, 2002.

E.Charniak, "A maximum-entropy inspired parse", In Proceedings of the *First Meeting of the North American Chapter of the Association for Computational Linguistics*, Seattle, WA. 2000, pages 132–139.

H.Samin, S.Nisar and S.Sehrai, *Corpus Development*, BIT thesis, Department of Computer Science, University of Peshawar, Peshawar, Pakistan, 2006.

K.Kewei Tu and V.Vasant Honavar, "Unsupervised Learning of Probabilistic Context-Free Grammar using Iterative Biclustering", In Proceedings of *9th International Colloquium on Grammatical Inference*, St Malo, Brittany, France, September 22-24, 2008.

M.A.Khan, F.T.Zuhra, "Pashto Chart parser", unpublished, Department of Computer Science, University of Peshawar, 2010.

M.Collins, *Head-driven statistical models for natural language parsing*, Ph.D. Dissertation, 1999.

M.Humayoun, H.Hammarström and A.Ranta, "Urdu Morphology, Orthography and Lexicon Extraction", CAASL-2, The Second Workshop on *Computational Approaches to Arabic Script-based Languages*, LSA 2007 Linguistic Institute, Stanford University, 2007.

R.Schmidt, "Urdu: an essential grammar", Rout-ledge, London, UK, 1999.

S.Petrov, L.Barrett, R.Thibaux and D.Dan Klein, "Learning accurate, compact and interpretable tree annotation", Proceedings of ACL, 2006.

Q.Abbas, N.Karamat and S.Niazi, "Development of Tree-bank Based Probabilistic Grammar for Urdu Language", International Journal of Electrical & Computer Sciences IJECS Vol: 9 No: 9, 2009.

W.Anwar, X.Wang, Luli and X.Wang, "Hidden Markov Model Based Part of Speech Tagger for Urdu", Information Technology, Vol.6, 2007, pages 1190-1198.