Relation Extraction for Matrix(type) entities in Introductory programing problems

Himanshu Shukla, Kumar Gaurav
Computer Science and Engineering Department
Indian Institute of Technology Kanpur
{hshukla,krgaurav}@cse.iitk.ac.in

Abstract

Relation Extraction has been a very important field since the start of Natural Language Processing. Relation Extraction has been studied in various fields eg.: bio-informatics, organization-employee relations, etc. We study the relation extraction for the mathematical(matrix type) entities in introductory programing problems. We use technique of Statistical Machine Translation for this by defining a bridging language(restricted domain language) for natural English.

1. Introduction

RELATIONS are important features inside any text and these things become more important when we enter into programing domains specially for the beginners who faces a lot of problem while programing. We define relations for the matrix type mathematical entities do extraction for the same in this project however the method that we use can be extended for other type of programing problems. Relation for the matrix type entities in introductory programing problems are the their attributes (size, contains, type, etc.) and operations (sum, sort, rows, etc). We define a domain specific language which we call as Bridging language which is similar to the bridging language defined in Pankaj et. al. 2014. We use Statistical Machine Translation tool MosesDecoder[3] and GIZA++[4] for mapping the natural English statements to metalanguage. The metalanguage is a formal grammar based language which can be parsed using Lex-Yacc compiler software. The Yacc software is also used for analyzing the semantics with reductions. The same is used for resolving anaphoras present in the metalanguage.

2. Related Work

A lot research has been done in the field of Relation Extraction. There has been significant work by Alessandro Moschitti over semantic mapping between natural language sql-queries[1]. Kernel methods have been widely used in the process of (employee-organization) relation extraction in normal english. Also a significant work has been done for relation extraction in the field of bio-informatics whose example is Stanford NLP.

3. Rejection of Kernel Methods

Work of Alessandro Moschitti was closest to our problem because it mapped the natural language questions to sql queries which is a restricted domain language(related to geolocation queries). We rejected this work because this method has got much less accuracy even small domain as mentioned by the the paper Moschitti et. al.[2].

4. Theory

Statistical Machine Translation is a technique of translating the sentence of one language called source language to the target language. This uses information theory at its base.

- The sentence is translated as according to the probability distribution $P(e|f)$ where e represents the event that sentence translation is e given that the foreign language sentence is f.
- Finding the best translation $\tilde{e}$ is done by picking up the one that gives the highest probability:

$$\tilde{e} = arg\max_{e \in e^*} p(e|f) = arg\max_{e \in e^*} p(f|e)p(e)$$

$P(F|E)$ Translation model
$P(E)$ Language model

- The translation is done sentence by sentence by using approximated smoothed n-gram language models. STMs are of three types: Word Based, Phrase Based and Syntax Based.

The Word Based alignment model was one of the initially used SMT's and we start with GIZA++ which is the word based alignment model to generate the probability mappings of different word of the English language to the bridging language. GIZA++ uses the HMM for generation of mappings.

5. Methodology

The methodology that we apply is as follows:

- Mapping to the bridge-language : In this we are presently using the technique of STM to map the natural language statement to bridge-language statement. We are using Phrase based Model available with MOSES-DECODER tool which uses GIZA++ for the word alignment.
- Semantics Analyzer: After the translation is done the semantics analyzer parses the metalanguage obtained to extract the relations between the entities, attributes and operations mentioned in the statement. Section 6 describes the working more explicitly.

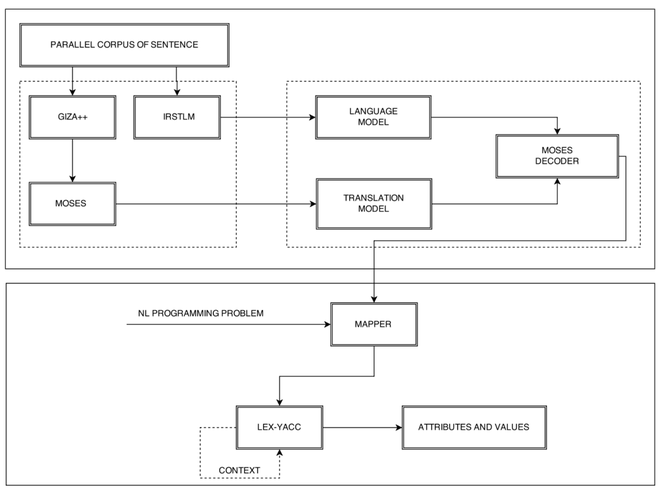

Figure 1: *Flowchart representing the basic work flow of the system.*

6. Semantics Analyzer

After the sentence has been aligned and mapped to the metalanguage, the control is passed over to a LALR parser generated using Lex-Yacc. The parser parses the obtained sentences using the rules mentioned in the parser and the parser derives the attributes present in the sentences simultaneously. Depending on the rule getting reduced within the grammar the parser assigns and maps values and attributes mentioned in the sentence with the corresponding entities(here matrices).

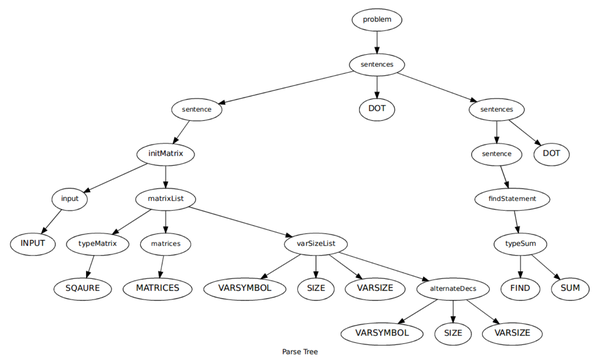

Figure 2: *Parse tree generated by running parser on a given problem statement in metalanguage.*

Another thing which happens during the reductions is Anaphora resolutions and removal. The system maintains the current sentence "context"[5] in a map containing all the entities whose values and attributes(if any) has been found till the previous sentence. The parser also maintains a "diff"[5] set of entities which are mentioned in the new sentence but the sentence itself has not been reduced yet. For example, consider a problem in metalanguage below:
*matrix A size 4*3.find determinant given matrix.*
In the above case, in general the parser would have not known the relation between words "A" and "given" unless rewritten as "A". But using the context sets the parser when encountered with "given" immediately looks back in the context set and assigns "A" as a possible candidate for the assignment "find determinant".

7. Corpus

Corpus contains the programing problems in english language and their bridge-language counterparts. We have made two files *corpus.en* and *corpus.me* that contain english and bridge-language statements respectively. Building of corpus has been done manually and the corpus presently contains 130 problem statements. Table represents a look of the corpus.en and the corresponding corpus.me.

Problem Statements in Natural Language	Problem Statement in Bridge-language
you are given a matrix of integer numbers with n rows and m columns	matrix size n rows m columns.
consider a matrix m of integers	integer matrix m.
find the inverse of the matrix.	find inverse matrix.
check if the matrices are identical or not.	check equality matrices
print a 2d matrix in a spiral form.	display matrix spiral.

8. Future Prospects

- Currently we are having a short corpus of about 130 natural language sentences so we are using only the word based SMT i.e GIZA++, in future we hope to extend the corpus and generate the translation using MosesDecoder and also use Phrasal the toolkit by the Stanford NLP group.
- Further we will include the anaphora handling in the grammar of the bridge-language.
- We also hope to extend this bridge-language for other matrix entities like arrays, strings, vectors, etc.

References

[1] Alessandra Giordani and Alessandro Moschitti. Semantic mapping between natural language questions and sql queries via syntactic pairing. In *Natural Language Processing and Information Systems*, pages 207–221. Springer, 2010.

[2] Alessandra Giordani and Alessandro Moschitti. Generating sql queries using natural language syntactic dependencies and metadata. In *Natural Language Processing and Information Systems*, pages 164–170. Springer, 2012.

[3] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, et al. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions*, pages 177–180. Association for Computational Linguistics, 2007.

[4] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[5] Kewalramani Pankaj Prateek, Jeetesh Mangwani, Amey Karkare, Sumit Gulwani, and Amitabha Mukerjee. Anaphora without syntax-a multi-lingual approach for geometry constructions. 2014.