St. Vincent Pallotti College of Engineering & Technology

Department of Information technology

Academic Year 2020-2021

Major Project Synopsis

Project Title: **Word Phrase Alignment**

Group No.: **14**

Name of Industry: International Institute of Information Technology, Hyderabad

Name of Industry Mentor: Dr. Soma Paul
Dr. Vineet Chaitanya

Name of Guide: Prof. Pooja Walke

Name of Alumni Mentor: Ms. CBS Manaswini

Students Name: **1. Aditya Dale**
2. Gaurav Bhisikar
3. Sagar Malik

ABSTRACT

The topic for our major project is “Word-Phrase Alignment”. The main objective of this project is to develop software tools and a scheme for building word or phrase aligned corpus of English and Hindi optimally using existing open resources. NMT models that are used to translate the English language to Hind language train on corpus which is on sentence level so during training they learn to translate on sentence level, as a result the current NMT models are not able to generate correct alignment of translated Hindi sentence. So this project focuses on identifying the relation between words of both English and Hindi sentences and to develop a aligned parallel corpus on word/phrase level.

Word alignment is the natural language processing task of identifying translation relationships among the words (or more rarely multiword units) in a bitext, resulting in a bipartite graph between the two sides of the bitext, with an arc between two words if and only if they are translations of one another. Word alignment is typically done after sentence alignment has already identified pairs of sentences that are translations of one another.

Bitext word alignment is an important supporting task for most methods of statistical machine translation. The parameters of statistical machine translation models are typically estimated by observing word-aligned bitexts, and conversely, automatic word alignment is typically done by choosing that alignment which best fits a statistical machine translation model. The circular application of these two ideas results in an instance of the expectation-maximization algorithm.

HARDWARE & SOFTWARE REQUIREMENTS

System Requirements:

1. Intel Core i3 or above
2. 4 GB RAM or more
3. 100 GB Storage or more
4. GPU- Integrated or Dedicated(optional)

Software Requirements:

1. **Ubuntu OS:** Ubuntu is a Linux distribution based on Debian mostly composed of free and open-source software. Ubuntu is officially released in three editions: Desktop, Server, and Core for the Internet of things devices and robots. All the editions can run on the computer alone, or in a virtual machine. Ubuntu is a popular operating system for cloud computing, with support for OpenStack. Ubuntu's default desktop, as of version 17.10, is GNOME.

2. **Python 3:** Python is an interpreted, high-level, and general-purpose programming language. Created by Guido van Rossum and first released in 1991, Python's design philosophy emphasizes code readability with its notable use of significant whitespace. Its language constructs and object-oriented approach aim to help programmers write clear, logical code for small and large-scale projects.

   Python 3.0, released in 2008, was a major revision of the language that is not completely backward-compatible, and much Python 2 code does not run unmodified on Python3.

3. **UCCA:** Universal Conceptual Cognitive Annotation (UCCA) is a novel semantic approach to grammatical representation. It was developed in the Computational Linguistics Lab of the Hebrew University by Omri Abend and Ari Rappoport.

   The central idea of the project is to analyze and annotate natural languages using purely semantic categories and structure (a graph). Syntactic categories and structure are not part of the manual annotation and are ideally learned implicitly by the parsers.

4. **ACE Parser:** ACE is an efficient processor for DELPH-IN HPSG grammars, with support for most of the modern features these grammars require, including

   Parsing with:

   a. REPP support
   b. Built-in part-of-speech tagging and unknown word handling
   c. Optional call-out to TNT tagger
   d. Token mapping and lexical filtering
   e. Idiom filtering

LITERATURE SURVEY

In paper [1] the authors Liang Tian, Derek F. Wong, Lidia S. Chao, Francisco Oliveira. This paper discusses the aspect of providing a formula to describe the relationship between word alignments, phrase table, and machine translation performance. The author of the papers also focuses on formulating such a relationship for estimating the size of extracted phrase pairs given one or more word alignment points.

In paper [2] the authors conducted a survey on Performance Improvement of Word Phrase Alignment in English to Hindi Translation System. In this paper the authors reviewed and discussed about different types of techniques of word alignment for English to Hindi translation system. This paper uses a combined technique to progress the implementation of word alignment for English-Hindi language pair with limited resources.

In paper [3] the author Santos, André presents the field of parallel corpora alignment. Its historical background and first development initiatives were described, and the alignment process was detailed step-by-step.

- **Parallel Text:** A parallel text is a set formed by a text and its translation (in which case it is called a bitext) or translations.

- **Parallel text alignment:** Parallel text alignment is the task of identifying correspondences between blocks or tokens in each halve of a bitext.

- **Alignment Details:** The alignment of texts may be performed at several levels of granularity. Usually, when the objective is to achieve low-level

alignment, like word alignment, higher-level alignment is performed first, which allows obtaining improved results.

REFERENCES

[1] Liang Tian, Derek F. Wong, Lidia S. Chao, Francisco Oliveira, "A Relationship: Word Alignment, Phrase Table, and Translation Quality", The Scientific World Journal, vol. 2014, Article ID 438106, 13 pages, 2014.

[2]A Survey Paper on Performance Improvement of Word Alignment in English to Hindi Translation System by Kamala Kant Yadav and Dr. Umesh Chandra Jaiswal.

[3] Santos, André. (2011). A survey on parallel corpora alignment.