Lit Review

Evan Chapple, Elizabeth Gabel, Emil Nadimanov, Tatiana Yudina

March 2021

1 Review of Literature

1.1 Chukchi

The language, Chukchi, is notable for being highly inflectional and morphologically rich. The morphemes that are found on Chukchi nominals, nouns, personal pronouns, indefinite pronouns, demonstrative pronouns, quantifier pronouns, and participles, give in depth information about the nominal itself. This information ranges from spatial relationships, semantic relations (cases), and others (Dunn, 1999). And there exists three categories into which all nominals fall to decide declension, animate, common, and high animate. Verbs also contain a lot of morphemes, that relay all sorts of information. Adding to a verb stem can add information on mood, tense, and aspect. It should also be noted that the specific inflection for a verb is directly modified by the semantic meaning associated with the verb in context. Another feature of Chukchi that leads to long words is a process referred to as incorporation, in which multiple lexical stems can be combined into one. This process can be performed in one of two ways, syntactic and lexical incorporation (Dunn, 1999).

1.2 Predictive Stuffs

While frequency of words is essential when determining the most likely next-typed word, another important aspect of predictive text is context-based analysis. In their 2004 paper *A Common Sense Approach to Predictive Text Entry*, Faaborg, Lieberman, and Stocky found that a commonsense approach to predictive text offered an additional layer to accurate prediction. The study evaluated the effectiveness of four approaches to predictive text on an email corpus: Language Frequency (5,000 most common words in English language), User Frequency (user's most frequent words), Recency (user's most recently typed), and Commonsense (a system which examined the semantic context of each word typed by the user and suggested words based on semantic matching). The study found that in most situations, the commonsense approach was on par with the other three conditions, but that it excelled in instances of low word repetition with increased word rarity (Faaborg et al., 2004). At the time, the greatest limitation on the application of the commonsense approach to predictive text was memory space, and issue which has been reduced with each new generation of technology since.

It is discussed in (Singh and Singh, 2019) how spellcheckers are developed and implemented for highly inflected languages. There are various techniques discussed due to the wide variety of languages and inflection. These techniques of processing natural language input, range from the most simple rule based system, to one that involves machine learning. One important innovation before this can even take effect is modifying the input string into digestable chunks that the computer can work upon. The original technique would be to take whole words and run analyses upon that data, having the dictionary include morphologically dense packages. This worked well enough for non-agglutinative languages or less morphology-rich languages but suffered greatly when applied to more highly inflected languages. This is where a technique referred to as Statistical Sadhi Splitter (SSS) came into play, which is discussed to a great extent in (Kuncham et al., 2015). The SSS takes input as a valid string in the language and then splits the input into valid substrings, usually meaningful morphemes, and then modifies those substrings into ones that can understood by the dictionary. The program can then interpret this final output and perform an operation there upon.

References

Dunn, M. J. (1999). *A Grammar of Chukchi: A thesis submitted for the degree of Doctor of Philosophy of Australian National University*. Australian National University.

Faaborg, A., Lieberman, H., and Stocky, T. (2004). A common sense approach to predictive text entry. *CHI EA '04: CHI '04 Extended Abstracts on Human Factors in Computing Systems*, pages 1163–1166.

Kuncham, P., Nelakuditi, K., Nallani, S., and Mamidi, R. (2015). Statistical sandhi splitter for agglutinative languages. *Lecture Notes in Computer Science*, pages 164–172.

Singh, S. and Singh, S. (2019). Systematic review of spell-checkers for highly inflectional languages. *Artificial Intelligence Review*.