Chapter 5

Computational Linguistic Distances and Big Data: Optimising the Speech Recognition Systems

Krishnaveer Abhishek Challa
Andhra University, India

Jawahar Annabattula
K. L. University, India

ABSTRACT

Linguistic distance has always been an inter-language issue, but English being a interwoven cluster of rhyming words, homophones, tenses etc., has turned linguistic distance into an intra-language unit to measure the similarity of sounds. In many theoretical and applied areas of computational linguistics i.e., Big Data, researchers operate with a notion of linguistic distance or, conversely, linguistic similarity has become the means to optimise speech recognition systems. The present research paper focuses on the mentioned lines as an attempt to turn the existing systems from delivering good performance to perfect performance, especially in the area of Big Data.

INTRODUCTION

Linguistic distance is how different one language or dialect is from another. Although there is no uniform approach to quantifying linguistic distance between languages, the concept is used in a variety of linguistic situations, such as learning additional languages, historical linguistics, language-based conflicts and the effects of language differences on trade. The proposed measures used for linguistic distance reflect varying understandings of the term itself. One approach is based on mutual intelligibility, i.e. the ability of speakers of one language to understand the other language. With this, the higher the linguistic distance, the lower is the level of mutual intelligibility (Chiswick & Miller, 2005).

DOI: 10.4018/978-1-5225-2947-7.ch005
The quest among linguists for a scalar measure of linguistic distance has been in vain. There is no yardstick for measuring distances between or among languages, as there is for the geographic distance between countries (e.g., miles). This arises because of the complexity of languages, which differ by vocabulary, grammar, syntax, written form, etc. The distance between two languages may also depend on whether it is in the written or spoken form. For example, the written form of Chinese does not vary among the regions of China, but the spoken languages differ sharply. Alternatively, two languages that may be close in the spoken form may differ more sharply in the written form (for example, if they use different alphabets, as in the case of German and Yiddish).

Perhaps the way to address the distance between languages is not through language trees which trace the evolution of languages, but by asking a simpler question: How difficult is it for individuals who know language A to learn languages B1 through Bi, where there are ‘i’ other languages. If it is more difficult to learn language B1, than it is to learn language B2, it can be said that language B1 is more “distant” from A than language B2. Language B3 may be as difficult to learn as is language B1 for a language A speaker, but that does not mean that language B3 is close to language B1. Indeed, it may be further from B1 than it is from A.

Linguistic distance is a concept that seeks to measure the degree of difference between two languages. Since the linguistic distances between languages are as different and variable as the languages themselves, such a concept cannot be accurately applied in a scientifically precise manner. This concept is important due to the increase in globalization, which has led to international trade between business concerns from different countries with different languages and dialects. It is also relevant as a tool to measure the ability of immigrants learning a new language that is different from their mother tongue. This is because the more removed one language is from another; the more difficult it will be for the immigrant to adapt to the new language. This also has big applications in the field of Computational Linguistics (CL).

Linguistic distance can be measured by measuring the mutual intelligibility of the language to the speakers. Mutual intelligibility determines how easy or difficult it will be for the speakers to grasp the fundamentals of the new language. This may be facilitated by the sharing of some common words or the similarity in the arrangement of grammatical and lexical forms. For instance, different territories or countries may speak the same basic language with only some minor or major differences in intonation, meaning of words, and the application of the language in general.

American and British English, for example, are mostly related with only a few easily surmountable variations. The linguistic distance between the methods of speaking the language is very small. On the other hand, Irish brogue and Cockney accent might prove to be a greater challenge for an American listener even though they are still variations of the same language. For these, the linguistic distance is more than that of British English. Even at that, learning to understand and speak these versions of the English language would not be as challenging as learning to speak Russian, since both versions are more related to American English and have a higher measure of mutual intelligibility.

The ability to easily overcome the linguistic distance between two languages is easier for children below the age of seven than it is for adults. This is due to the fact that there is a threshold in which children are still open to absorbing the fundamentals of speech patterns without encumbrance from other learned languages. Children below seven are still able to assimilate the fundamentals of a new language, master it, and speak the language without an accent better than adults. This effect can be observed in new immigrants to a country with a different language. The adults may eventually learn the new language, but the possibility of retaining accents carried over from the mother tongue is very high, especially if the linguistic distance is high. This theory can also be applied in Computation.
Related Content

EMG-Based Mobile Assessment System for Neck and Shoulder Fatigue
[www.igi-global.com/article/emg-based-mobile-assessment-system-for-neck-and-shoulder-fatigue/204447?camid=4v1a](www.igi-global.com/article/emg-based-mobile-assessment-system-for-neck-and-shoulder-fatigue/204447?camid=4v1a)

Vignette Methodology: An Illustration From Conflict Research
Swati Alok (2019). *Qualitative Techniques for Workplace Data Analysis* (pp. 117-143).
[www.igi-global.com/chapter/vignette-methodology/207794?camid=4v1a](www.igi-global.com/chapter/vignette-methodology/207794?camid=4v1a)

Application of Geographical Information System and Interactive Data Visualization in Healthcare Decision Making

Conceptual View on Healthcare Digitalization: An Extended Thematic Analysis
[www.igi-global.com/article/conceptual-view-on-healthcare-digitalization/197440?camid=4v1a](www.igi-global.com/article/conceptual-view-on-healthcare-digitalization/197440?camid=4v1a)