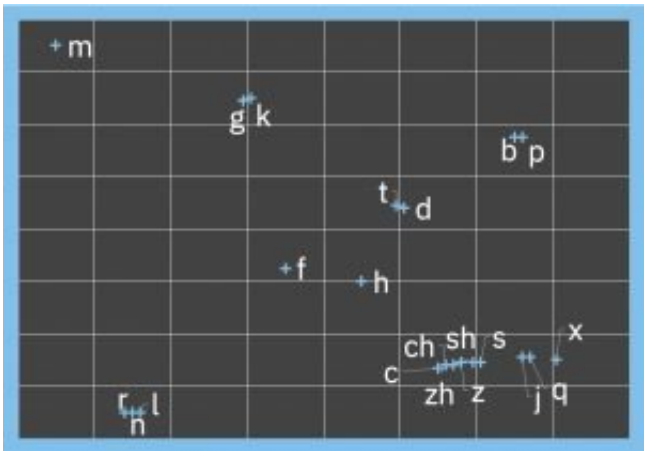
Learning Chinese-specific encoding for phonetic similarity

November 6 2018, by Marina Danilevsky



Visualization representing the phonetic encoding of Pinyin initials. Credit: IBM

Performing the mental gymnastics of making the phoenetic distinction between words and phrases such as "I'm hear" to "I'm here" or "I can't so but tons" to "I can't sew buttons," is familiar to anyone who has encountered autocorrected text messages, punny social media posts and the like. Although at first glance it may seem that phonetic similarity can only be quantified for audible words, this problem is often present in purely textual spaces.

AI approaches for parsing and understanding text require clean input, which in turn implies a necessary amount of pre-processing of raw data. Incorrect homophones and synophones, whether used in error or in jest,

must be corrected just like as any other form of spelling or grammar mistake. In the example above, accurately transforming the words "hear" and "so" to their phonetically similar correct counterparts requires a robust representation of phonetic similarity between word pairs.

Most algorithms for phonetic similarity are motivated by English use cases, and designed for Indo-European languages. However, many languages, such as Chinese, have a different phonetic structure. The speech sound of a Chinese character is represented by a single syllable in Pinyin, the official Romanization system of Chinese. A Pinyin syllable consists of: an (optional) initial (such as 'b', 'zh', or 'x'), a final (such as 'a', 'ou', 'wai', or 'yuan') and tone (of which there are five). Mapping these speech sounds to English phonemes results in a fairly inaccurate representation, and using Indo-European phonetic similarity algorithms further compounds the problem. For example, two well-known algorithms, Soundex and Double Metaphone, index consonants while ignoring vowels (and have no concept of tones).

Pinyin

As a Pinyin syllable represents an average of seven different Chinese characters, the preponderance of homophones is even greater than in English. Meanwhile, the use of Pinyin for text creation is extremely prevalent in mobile and chat applications, both when using speech-to-text and when typing directly, as it is more practical to input a Pinyin syllable and select the intended character. As a result, phonetic-based input mistakes are extremely common, highlighting the need for a very accurate phonetic similarity algorithm that can be relied on to remedy errors.

Motivated by this use case, which generalizes to many other languages that do not easily fit the phonetic mold of English, we developed an approach for learning an n-dimensional phonetic encoding for Chinese,

An important characteristic of Pinyin is that the three components of a syllable (initial, final and tone) should be considered and compared independently. For example, the phonetic similarity of the finals "ie" and "ue" is identical in the Pinyin pairs {"xie2," "xue2"} and {"lie2," "lue2"}, in spite of the varying initials. Thus, the similarity of a pair of Pinyin syllables is an aggregation of the similarities between their initials, finals, and tones.

However, artificially constraining the encoding space to a low dimension (e.g., indexing every initial to a single categorical, or even numerical value) limits the accuracy of capturing the phonetic variations. The correct, data-driven approach is therefore to organically learn an encoding of appropriate dimensionality. The learning model derives accurate encodings by jointly considering Pinyin linguistic characteristics, such as place of articulation and pronunciation methods, as well as high quality annotated training data sets.

Demonstrating a 7.5X Improvement Over Existing Phonetic Similarity Approaches

The learned encodings can therefore be used to, for example, accept a word as input and return a ranked list of phonetically similar words (ranked by decreasing phonetic similarity). Ranking is important because downstream applications will not scale to consider a large number of substitute candidates for each word, especially when running in real time. As a real world example, we evaluated our approach for generating a ranked list of candidates for each of 350 Chinese words taken from a social media dataset, and demonstrated a 7.5X improvement over existing phonetic similarity approaches.

We hope that the improvements yielded by this work for representing language-specific phonetic similarity contributes to the quality of

numerous multilingual natural language processing applications. This work, part of the IBM Research SystemT project, was recently presented at the 2018 SIGNLL Conference on Computational Natural Language Learning, and the pre-trained Chinese model is available for researchers to use as a resource in building chatbots, messaging apps, spellcheckers and any other relevant applications.

More information: DIMSIM: An Accurate Chinese Phonetic Similarity Algorithm based on Learned High Dimensional Encoding: aclweb.org/anthology/K18-1043

Provided by IBM