September 19, 2019

Phonetic training of spelling models

Sanjit Jhala

Pratibha Prajapati

Bing Liu

Grant Wang

Phonetic training of spelling models

ABSTRACT

This disclosure describes techniques for phonetic training of spelling models for use for spell correction. Phonetic canonicalization is utilized during the training of the spelling model to generate rules that map similar sounding words and portions of words to each other. Phonetic normalization is utilized to reduce the space of phonetic representations. A combination of a textual edit distance and a phonetic edit distance is utilized to score corrected alternatives for a word. The minimum of the textual edit distance and the phonetic edit distance is used as the noisy channel edit distance. Phonetic canonicalization can also be utilized during runtime in addition to its utilization in the training of the spelling model.

KEYWORDS

- Spell correction
- Phonetic canonicalization
- Phonetic normalization
- Phonetic distance
- Edit distance
- Multilingual
- Phonetic boost
- Noisy channel

BACKGROUND

Spell correction is a commonly used feature in many applications such as search engines (query spell correction), email applications, word processors, etc. Spell correction typically

utilizes spelling models that are trained on pairs of input and corrected queries known as training instances.

The training instances are generated using a language model that assigns probabilities to sequences of words, a noisy channel edit distance (e.g. textual edit distance) that is a measure of a difference (distance) between a misspelled word and an intended word, and word replacement rules that are generated by various canonicalization methods. Word replacement rules are used to generate spell correction candidates for a user query, while the language model and the noisy channel edit distance are used to score the spell correction candidates relative to the user query.

However, users often misspell queries phonetically. For example, users who may not have formally learnt the English language, but who are aware of the English language and alphabet, and sounds associated with different letters of the alphabet, often spell their query based on pronunciation (phonetically) without knowing the correct spelling. For example, ‘pattern’ may be phonetically misspelled as ‘petan,’ ‘train’ as ‘teran,’ ‘table’ as ‘tebal,’ etc. Phonetic misspellings such as these have a pronunciation similar to that of the correct word, even if spelled differently.

Spell correction of phonetic misspellings using traditional spelling models poses challenges due to low scores being assigned to the correct query in the language model and a high textual edit distance between the misspelled query and the correct query.

DESCRIPTION

This disclosure describes correction of phonetic misspellings in training instances utilized in the training of spelling models. Phonetically trained spelling models can provide superior spell correction performance in the correction of phonetic misspellings.

Per techniques of this disclosure, phonetic canonicalization is utilized during the training of the spelling model to generate rules that map similar sounding words and portions of words (for example, n-grams) to each other. For example, ‘petan’ may be mapped to ‘pattern’ and ‘prjent’ may be mapped to ‘present’ due to their similar pronunciation. The different phonetic misspellings of a word are all mapped to the same normalized form to enable scoring. Phonetic normalization is utilized to reduce a space of phonetic representations. In addition, the noisy channel edit distance between a word and its corrected alternative is determined based on a combination of a textual edit distance and a phonetic edit distance.

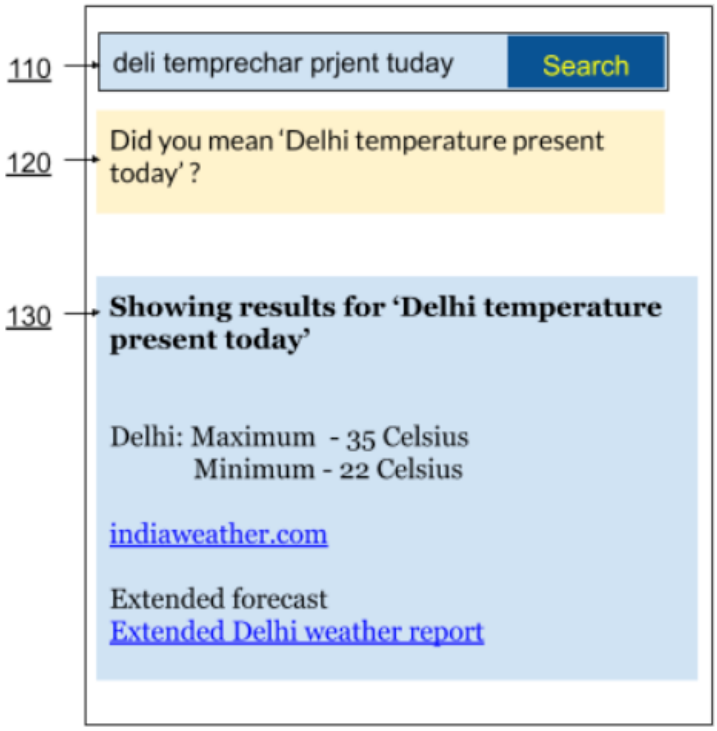

Fig. 1: Spell corrections using a phonetically trained model

Fig. 1 illustrates the use of a phonetically trained spelling model for spell correction. A user that intends to obtain information about temperatures in Delhi provides a query (110) ‘deli temprechar prjent today’ to a search engine. The search engine utilizes the phonetically trained

spelling model to determine (120) that the query intended by the user was ‘Delhi temperature present today.’ Based on the corrected query (as determined by the spelling model), the search engine provides results (130) based on the intended user query. Phonetically trained spelling models can thus provide superior performance in use cases involving users of a language without a formal instruction in the language.

In some implementations, the minimum of the textual edit distance and the phonetic edit distance is used as the noisy channel edit distance during training. In some implementations, phonetic canonicalization is utilized at runtime in addition to its utilization in the training of the spelling model. The described techniques can provide better corrections than spell correction systems that use edit distance and rule-based sound similarity, and can cover a large set of languages, with a wide variety of phonetic misspellings. The techniques can be used in any spell correction context, e.g., web search, document editing, etc.

CONCLUSION

This disclosure describes techniques for phonetic training of spelling models for use for spell correction. Phonetic canonicalization is utilized during the training of the spelling model to generate rules that map similar sounding words and portions of words to each other. Phonetic normalization is utilized to reduce the space of phonetic representations. A combination of a textual edit distance and a phonetic edit distance is utilized to score corrected alternatives for a word. The minimum of the textual edit distance and the phonetic edit distance is used as the noisy channel edit distance. Phonetic canonicalization can also be utilized during runtime in addition to its utilization in the training of the spelling model.