Predictive Text Entry of Agglutinative Languages using Morphological Segmentation and Phonological Restrictions

...

...

Abstract. Linguistic models for predictive text entry on ambiguous keyboards typically rely on large dictionaries including word frequenies, which are used to disambiguate between words matching user input. This approach is insufficient for heavily agglutinative languages, like Finnish or Turkish, where morphological phenomena such as inflection and compounding increase the rate of out-of-vocabulary words. We propose a method for text entry, which circumvents the problem of out-of-vocabulary words, by replacing the dictionary with a Markov chain on morpheme sequences constructed from morphologically segmented training data. The Markov chain is combined with a third order hidden Markov model (HMM) mapping key sequences to letter sequences. Additionally we use rules, which enforce phonotactic restrictions such as vowel harmony in Finnish. We evaluate our method by constructing text entry systems for Finnish and Turkish mobile phone keypads. We compare the Turkish text entry system with an existing system, which is based on an HMM of letter sequences [6] and show that we achieve superior results measured by the keystrokes per character ratio (KPC) [3]. We also compare the Finnish text entry system to an existing system, which utilizes a morphological analyzer combined with a colloquial dictionary [4] and show that we achieve superior KPC. For segmenting the training data, we use Morfessor, a system for unsupervised morphological segmentation [1]. For constructing the probabilistic models needed for the text entry systems, we use tools for POS tagging from the HFST interface [5], which is an open-source interface for weighted finite-state calculus. We also utilize an open-source two-level phonology rule compiler, hfst-twolc, for implementing the vowel harmony rules needed for text entry of Finnish [2].

1 Introduction

2 Earlier Approaches to Predictive Text Entry

3 A Probabilistic Model of Word Structure

3.1 An Hidden Markov Model for Predicting Letter Sequences from Key Sequences

3.2 A Markov Chain of Morphs

3.3 Phonological Constraints

3.4 Combining Models using Weighted Finite-State Calculus

4 Training Materials and Test Materials for Finnish and Turkish

4.1 Finnish

4.2 Turkish

5 Evaluation

Many factors influence the efficiency of mobile phone text entry in practice. E.g. the general user interface design of the phone and specifically the design of the keyboard have a large impact. Nevertheless such factors are in some sense isolated from the predicitve text entry algorithm itself, which makes it plausible to evaluate the algorithms in isolation from the rest of the user interface of mobile phones. In this paper we use the keystrokes per character (KPC) ratio [3] for measuring the efficiency of text entry. It measures the average number of keystrokes required to input one letter in a text message.

5.1 The Keystrokes Per Characters Ratio

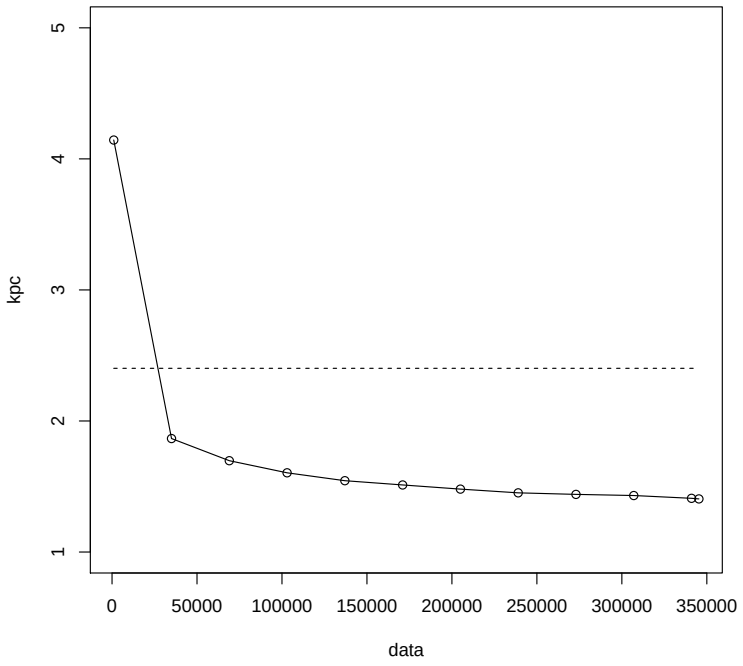
6 Discussion

7 Conclusions

8 Acknowledgments

References

1. Creutz, M., Lagus, K.: Unsupervised models for morpheme segmentation and morphology learning. ACM Transactions on Speech and Language Processing 4(1) (January 2007)



2. Lindén, K., Axelson, E., Hardwick, S., Silfverberg, M., Pirinen, T.: Hfst–framework for compiling and applying morphologies —, 67–85 (2011)
3. MacKenzie, I.S.: Kspc (keystrokes per character) as a characteristic of text entry techniques. In: Proceedings of the Fourth International Symposium on Human Computer Interaction with Mobile Devices. pp. 195–210. Springer-Verlag (2002)
4. Silfverberg, M., Hyvärinen, M., Pirinen, T.: Improving predictive entry of finnish text messages using irc logs. In: Proceedings of Computational Linguistics and Applications. Jachranka, Poland (2011)
5. Silfverberg, M., Lindén, K.: Combining statistical models for pos tagging using finite-state calculus. In: Proceedings of the 18th Conference on Computational Linguistics, NODALIDA 2011. pp. 183–190 (2011)
6. Tantuğ, A.C.: A probabilistic mobile text entry system for agglutinative languages. IEEE Transactions on Consumer Electronics 56(4) (2010)