Int. J. Human-Computer Studies 113 (2018) 15–31

Contents lists available at ScienceDirect

International Journal of Human-Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Multilingual phrase sampling for text entry evaluations

Marc Franco-Salvador[a], Luis A. Leiva[b,*]

[a] Symanto Research, Pretzfelder Strasse 15, Nuremberg DE-90425, Germany
[b] Sciling, SL, Camí a la Mar, 75, Valencia 46120, Spain

ARTICLE INFO

Keywords:
Phrases
Sentences
Sampling
Multilingualism
Memorability
Representativeness
Semantics

ABSTRACT

Text entry evaluations are typically conducted with English-only phrase sets. This calls into question the validity of the results when conducting evaluations with non-native English speakers. Automated phrase sampling methods alleviate this problem, however they are difficult to use in practice and do not take into account language semantics, which is an important attribute to optimize. To achieve this goal, we present KAPS, a phrase sampling method that uses the BabelNet multilingual semantic network as a common knowledge resource, aimed at both *standardizing* and *simplifying* the sampling procedure to a great extent. We analyze our method from several perspectives, namely the effect of sampled phrases on user's foreign language proficiency, phrase set memorability and representativeness, and semantic coverage. We also conduct a large-scale evaluation involving native speakers of 10 different languages. Overall, we show that our method is an important step toward and provides unprecedented insight into multilingual text entry evaluations.

1. Introduction

Text entry as a research discipline is increasingly attracting the interest of many researchers. By way of example, as of January 2017 the query (“text entry” “text input”) returns 647 results in the ACM digital library, and we can see that the number of publications has doubled each lustrum in the last 15 years. These figures put forward the fact that text entry is a rapidly growing community.

Some authors argue that text entry research has seen a revival in recent years due to the advent of mobile devices. Indeed, academic and industry researchers have been working on text entry since the emergence of handheld technologies (Dunlop and Masters, 2009; Wobbrock and Myers, 2006). Eventually, as any other interaction technique, text entry methods need to be evaluated. However, it is well-known that the outcome of text entry experiments is affected by the text users enter (Mackenzie and Felzer, 2010).

1.1. Text entry evaluations

Typically, in text entry experiments participants are prompted with phrases (short sentences) that must be entered as quickly and accurately as possible. Phrases can be full sentences or sentence fragments such as greetings, idioms, or quotations. An alternative option to *transcription* (i.e., copying pre-selected text) is *composition* (i.e., generating new text). This is considered more ecologically valid when evaluating text entry techniques (Zhai et al., 2005), since it mimics typical device usage. However text composition tasks are more difficult to control and measure. For example, since there is no reference text available, it is not possible to compute well-established measures of input error such as the character error rate. As a workaround, Vertanen and Kristensson (2014) described how human judging can provide a surrogate error rate measure that ensures participants are making good faith efforts to be accurate.

Overall, although it may seem more natural to have users enter free text and increase thus the external validity of the experiment (i.e., the extent to which the observed effect is generalizable), it is critical to make the text entry method the only independent variable in the experiment, and increase thus its internal validity (i.e., the extent to which the observed effect is due to the test conditions). Indeed, if users were asked to “type anything as fast as possible” they would introduce rather biased (maybe nonsensical) text. Hence, text entry researchers typically use pre-selected phrases, measuring the dependent variables (e.g., input speed or error rates) in a transcription task. This eliminates noise and facilitates the comparison of text input techniques across studies (Leiva and Sanchis-Trilles, 2014; Vertanen and Kristensson, 2011b).

In general, transcription tasks should prefer memorable stimuli (Leiva and Sanchis-Trilles, 2014; MacKenzie and Soukoreff, 2002; Vertanen and Kristensson, 2011b). This reduces participants' tendency to shift attention between the stimulus phrase and the text entry method. To ensure memorable stimuli, researchers often resort to using manually curated English-only phrase sets, which are typically small according to

* Corresponding author.
E-mail addresses: marc.franco@symanto.net (M. Franco-Salvador), name@sciling.com (L.A. Leiva).

https://doi.org/10.1016/j.ijhcs.2018.01.006
Received 3 February 2017; Received in revised form 19 January 2018; Accepted 22 January 2018
Available online 31 January 2018
1071-5819/

modern standards, or rely on sampling procedures that do not guarantee the internal validity of the experiment. In contrast, today text is entered into many different devices in many different languages, where text entry methods might perform very differently. This fact evidences the necessity of an adequate phrase sampling method, aimed at exploiting the huge amount of text corpora available in different languages.

Currently, most text entry experiments in HCI involving English users use either the phrase set released by MacKenzie and Soukoreff (2003) (MacKenzie dataset for short) or the EnronMobile dataset (Vertanen and Kristensson, 2011b). Both phrase sets have been proved to be adequate for conducting experiments with English participants. Furthermore, using these sets make it easier to reproduce different studies conducted by other researchers. The problem, however, is how to conduct text entry experiments with non-English users or in very specialized fields (e.g. a text entry method for a medical device, where technical vocabulary is commonplace), or simply when the participants speak different languages.g It is here where automated sampling methods like Ngram (Paek and Hsu, 2011) or MemRep (Leiva and Sanchis-Trilles, 2014) are valuable. On the one hand, the Ngrammethod approaches phrase sampling as a single-objective function, which only considered the representativeness of the phrases as the measure to optimize (to be described later). On the other hand, the MemRep sampling method approaches phrase sampling as a dual-objective function, incorporating both representativeness and memorability as the measures to optimize (also to be described later).

To the best of our knowledge, to date MemRep is the only automated method that provides adequate phrases for conducting text entry experiments in languages different from English. However, MemRep is cumbersome to use in practice, since it requires an additional large dataset to learn a language model, in addition to the input dataset from which phrases will be sampled. Moreover, MemRep does not take into account phrase semantics, which may result in confusing phrases like '`Send e.m.s.`' or '`100 is proposed for this category`'. We elaborate this discussion in the next sections.

1.2. Contributions

In this article, we present Kaps (acronym for Knowledge-Aided Phrase Sampling), an automated method for sampling phrase sets that uses knowledge graphs to select phrases for text entry seeking a balance among memorability, representativeness, and semantics. Our method is based on a multiple regression model over language-independent features, so that it can generalize to other languages. Kaps uses the BabelNet multilingual semantic network as a common resource, for *standardization* purposes, which also *simplifies* the evaluation procedure to a great extent. For example, contrary to MemRep , Kaps does not need a statistical analysis of an additional (large) corpus, just the dataset from which phrases will be drawn. An interesting property of our method is that, being data-driven, the sampled phrases are prototypical of the language or domain of interest.

We analyze Kaps from several perspectives, namely the effect of sampled phrases on user's foreign language proficiency, phrase set memorability and representativeness, and semantic coverage. We also conduct a large-scale evaluation involving 200 native speakers in 10 different languages. Overall, we show that our method is an important step toward and provides unprecedented insight into multilingual text entry evaluations. Finally, we make our software and data (ready-made phrase sets) publicly available so that others can build upon our work.

2. Related work

The choice of phrase set has been extensively discussed in the text entry literature. In the past, researchers used ad-hoc text sources for their experiments, such as sentences drawn from a Western novel (Karat et al., 1999), quotations from Unix's `fortune` program (Isokoski and Raisamo, 2000), news snippets (Zhai et al., 2002), street addresses (González et al., 2007), or passages from Sherlock Holmes (Tanaka-Ishii et al., 2003) and Alice in Wonderland (Vasiljevas et al., 2015). Using ad-hoc, proprietary text sources is often considered a bad practice because text entry studies could not be accurately reproduced. To help the situation, (MacKenzie and Soukoreff, 2003) released a phrase set consisting of 500 phrases that was also designed to contain easy to remember text. However, the MacKenzie phrase set mainly consists of short English idioms and clichés, and the memorability of the phrase set was never verified. Vertanen and Kristensson (2011b) processed the Enron email dataset and released the EnronMobile phrase set, including empirical data regarding sentence memorability.

Table 1
Survey of recent research on text entry involving user studies, according to the ACM digital library.

Language	No. Studies	Dataset Used		
		MacKenzie	EnronMobile	Custom
English	91	55	10	26
French	15	7[a]		8
Finnish	14	11[b]		3[c]
Portuguese	12	2[d]	1[d]	9
German	12	5	1	6
Korean	8	5[e]		3
Hindi	7	2		5
Chinese	6	1		5
Spanish	5	2		3[f]
Japanese	4	2		2
Dutch	3	1		2
Italian	2			2
Africans	2	1		1
Norwegian	1	1		
Bulgarian	1			1
Arabic	1	1		
Myanmar	1			1
Bengali	1			1
Total	185	96	12	78

[a] 2 were translated to French.
[b] 7 were translated to Finnish.
[c] 1 was left in English.
[d] 2 were translated to Portuguese.
[e] 2 were translated to Korean.
[f] 1 study used also the original MacKenzie dataset.

Both MacKenzie and EnronMobile are today the most popular phrase sets used in text entry experiments. Kristensson and Vertanen (2012) compared both phrase sets and found not much difference between them, although the actual differences are *conceptually* rather large. For example, EnronMobile is better suited to evaluating mobile text entry methods, as it contains genuine mobile emails. Other researchers have developed alternative phrase sets for specialist applications. For example, Kano et al. (2006) curated a phrase set for specific use with children and Vertanen and Kristensson (2011a) created a phrase set for Augmentative and Alternative Communication (AAC) by using messages suggested by AAC specialists.

2.1. Multilingual text entry

It has been argued that the choice of phrase set might not matter much as long as it is memorable and somewhat representative of the text users write (Vertanen and Kristensson, 2011b). However, current standard datasets for text entry are only available in English. Meanwhile, it is clear and obvious that "text entry" does not imply "English text entry" (MacKenzie and Soukoreff, 2002). Many text entry researchers are conducting user studies in many languages different from English; see Table 1.