Effects of including more words in text units on bilingual predictive dictionaries

Gábor Szabó, Sándor Juhász Dr.
Department of Automation and Applied Informatics
Budapest University of Technology and Economics
Budapest, Hungary
harry.bme@gmail.com, Juhasz.Sandor@aut.bme.hu

***Abstract*—In this paper we summarize the research in the field of giving predictive translation typing capabilities to computer-assisted translation tools, including the TransType project. Then we present our dictionary based method that uses simpler statistical techniques to extract possible text unit translations of a large, segmented, bilingual corpus. We investigated how using different types of text units (only words, bigrams or trigrams) affects the size and the goodness of the resulting dictionary. For measuring size of the dictionary both entry number and physical size are considered, and for measuring goodness we chose the ratio of number of keystrokes saved by translators to the length of the translation. Measurements were carried out for English-German and English-Hungarian language pairs.**

Keywords-predictive typing; dictionary based; multi-word text units; statistical translation extraction

I. INTRODUCTION

Computer systems have an important role in people's life: make it easier and more efficient. Predictive typing or automatic completion of text input is one tool to make human-computer interaction more efficient. It also helps to reduce the amount of human errors.

One outstanding example of this is using devices with limited text input capabilities. Predictive typing had an important role in speeding up text input on mobile phones with hard keys [7], by allowing to press a key on the phone only once for each character and disambiguating using a dictionary, i.e. trying to determine which word the user wanted to type. How and Kan suggested automatic word completion based on the keys typed and the previous word [4]. Masui investigated the text-input productivity gain for English and Japanese language by using automatic word completion on pen-based soft key input devices, based on a predictive dictionary that stores words with their context [5]. Approximate string matching based on spatial key layout was also presented in that work, to make the system tolerant for typos.

It is an interesting idea to make predictive typing available between different languages. It can be a powerful feature to make translator's work more productive inside a computer assisted translation tool, i.e. automatic completion is offered while the user is typing the translation of a text.

The groundbreaking research in predictive translation typing was the TransType project [2][3]. Researchers of that project used machine translation systems to produce possible translations that could complete the text being typed by the translator. Words or shorter sequence of words were offered as translation completion to the user in that project.

Research were continued in the TransType2 project, a successor of TransType. As a new approach, full sentence translation completions were offered to the translator. In each iteration she accepts some prefix of the completion, then types some other text, and the computer gives another full sentence completion. Alignment templates, phrase-based models and stochastic finite-state transducers were examined in TransType2 as translation models [1].

Machine translation engines are complex systems. This paper presents a simpler approach for predictive translation typing: it uses a dictionary to offer typing completions. In that sense, it is similar to the typing completion solutions used in mobile phones.

There are multiple aspects of creating such dictionary. One of them is how long text fragments are stored in the dictionary. The measurement presented in this paper compares three kinds of dictionaries: one that uses only one-word sequences, another that includes bigrams, and a thirds that includes bigrams and trigrams as well.

II. DICTIONARY BASED PREDICTIVE TRANSLATION TYPING

A. Concept

The concept of dictionary based prediction is to create a dictionary from some source once, and use it to give predictions many times later. A bilingual dictionary is created and used in the presented case. The source of the dictionary is a large, segment-aligned, bilingual corpus. The corpus is processed with statistical methods in order to find translations of text fragments.

Typing completions can be offered based on these text fragments: if a known text fragment is found in the source language sentence and the translator types something that is the beginning of a translation in the dictionary, then this known

translation can be shown as a possible completion. Text fragments are called text units in this paper. The exact definition of text unit is presented at the measurement's description, but they can be thought of like some words from a sentence that are linked together in a way, e.g. they form an expression or subsentence.

An entry in the dictionary is a source language text unit coupled with a list of possible translations. Each possible translation in an entry consists of a target language text unit and a score representing the probability of being a correct translation. The score is used when the typing completions are sorted and filtered. Filtering to the best n completions is necessary (n is around 5-6), otherwise too many completions would just hinder the translator's work. An example dictionary is presented in Table 1.

TABLE I. SAMPLE ENGLISH-GERMAN DICTIONARY FOR PREDICTIVE TYPING. EACH SOURCE LANGUAGE TEXT UNIT HAS A LIST OF POSSIBLE TRANSLATIONS ALONG THEIR SCORES

Source language text unit	Possible translations
I would	(möchte; 0.45), (möchte ich; 0.31)
new	(neuen; 0.55), (neue; 0.49)
Parliament	(Parlament; 0.75), (Parlaments, 0.40)
and I	(und ich; 0.54)
and	(und; 0.88), (der; 0.61), (die; 0.59), (in; 0.46), …

The structure of the dictionary implies that the method is asymmetrical, i.e. the dictionary can only be used for translation from the source language to the target language.

B. Dictionary creation

The dictionary is created in roughly three steps. Firstly, text units are collected from each side of the bilingual corpus individually (i.e. the source and target language). The "rare" text units are filtered out. Here, being "rare" is defined by the number of occurrences in the corpus. If the occurrence falls under a given threshold o_{min}, then the text unit is considered rare.

After that, the remaining text units of the source language text are paired with text units of the target language text, and their concomitant occurrences are counted (i.e. how often they appear together in a segment and its translation).

Then, a score is assigned to every text unit pair based on the individual and concomitant occurrences. If this score falls under a given threshold s_{min}, then the pair is omitted, otherwise it is added to the dictionary as a possible translation entry (along with the score). The score is assigned to a *(s; t)* source and target language text unit pair by applying the following formula:

$$score = 2 \cdot N_{s,t} / (N_s + N_t). \quad (1)$$

Here $N_{s,t}$ denotes the number of sentence pairs where s and t occurs together, and N_s and N_t denotes the number of sentences where s or t occur, respectively. Multiple occurrences in a sentence pair count as one.

This score is called the Dice coefficient comparator [6]. It falls in the [0; 1] interval and expresses the ratio of concomitant occurrences to the sum of the individual occurrences.

III. MEASUREMENT

There are basically two aspects of creating a dictionary: the exact definition of a text unit and the thresholds. These parameters can be exactly set before the measurement and they can alter the size of the resulting dictionary. The measurement presented in this paper uses a fix set of thresholds and three kind of text unit: words, words and bigrams, and words, bigrams and trigrams. Bigrams and trigrams cannot overlap subsentence boundaries.

The goal of the measurement is to investigate how using these text units affect the size and goodness of the dictionary. Size can be measured both in physical size (e.g. in megabytes) and in entry number.

The metric for the goodness of the dictionary is keystroke saving ratio, i.e. how much part of the typing can be saved if the translator uses the dictionary. It is automatically measured with a bilingual corpus. The source language sentences of the bilingual corpus are the sentences the hypothetical translator wants to translate, and the target language sentences are the translations the translator would like to type. When the simulated translator starts translating a new sentence, the sentence is split into text units, and these text units are looked up in the dictionary for possible translations. If the translator starts typing a new word, and it is the prefix of some possible translations, those possible translations are offered as completions. The best six completions is shown to the translator, and she chooses the longest appropriate completion. The number of keystrokes saved by the translator is: [length of completion] – [length of typed prefix] – [index of completion in the offered list].

Expanding the definition of text unit by including bigrams and trigrams obviously results at least the same dictionary size as the one using only words. This is because the text units, which are kept in the first step of dictionary creation using only words, are also kept including bigrams and trigrams. It is expected that some actual bigrams and trigrams survive the filtering of the first step and the scoring of the last step, meaning the dictionary size will definitely increase. The questions is whether the keystroke saving ratio will also increase.

The measurement program was written in C# language targeting .NET Framework 4.5. Measurements were carried out on a 64-bit Microsoft Windows 8.1 operating system, using an Intel Core i7-2630QM 2.00 GHz CPU with 8 GB RAM.

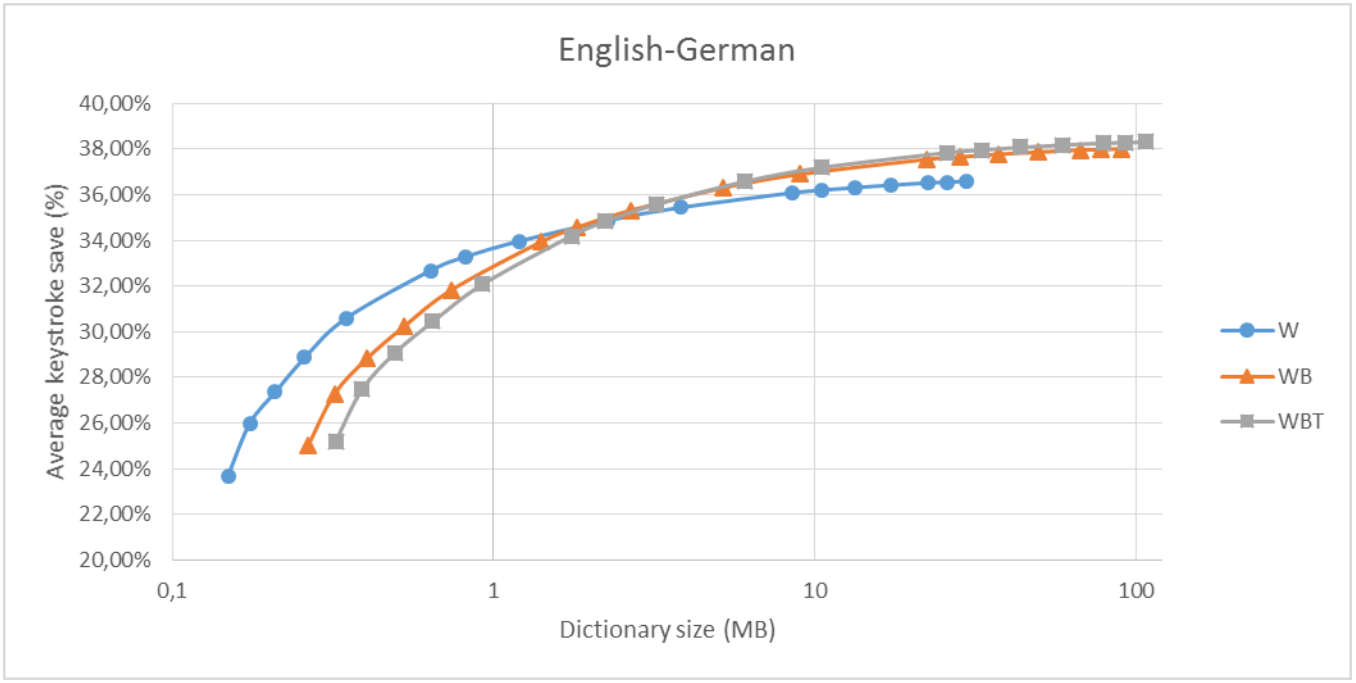

Figure 2. Keystroke saving results for English-German language pair, based on the physical size of the dictionary. For small dictionary sizes the word-only approach is more efficient, but including bigrams and trigrams result in greater keystroke saving beyond 3 megabytes.

The bilingual corpus were extracted from the proceedings of the European Parliament, downloaded from the OPUS project [8]. English-German and English-Hungarian language pairs were chosen. 200k sentence pairs were used for dictionary creation and 20k for keystroke saving measurement. The thresholds used: o_{min} = 1e-3, s_{min} = 0.3, 0.25, 0.2, 0.15, 0.1, 5e-2, 4e-2, 3e-2, 2e-2, 1.5e-2, 1e-2, 9e-3, 8e-3, 7e-3, 6e-3, 5.5e-3, 5e-3. Thresholds were chosen so that dictionary size is between a few hundred kilobytes and a hundred megabytes.

IV. Results

Measurement results can be seen on Fig. 1, Fig 2 and Fig 3. The dictionary sizes on the horizontal axes are displayed with a logarithmic scale for better readability. The letters on the

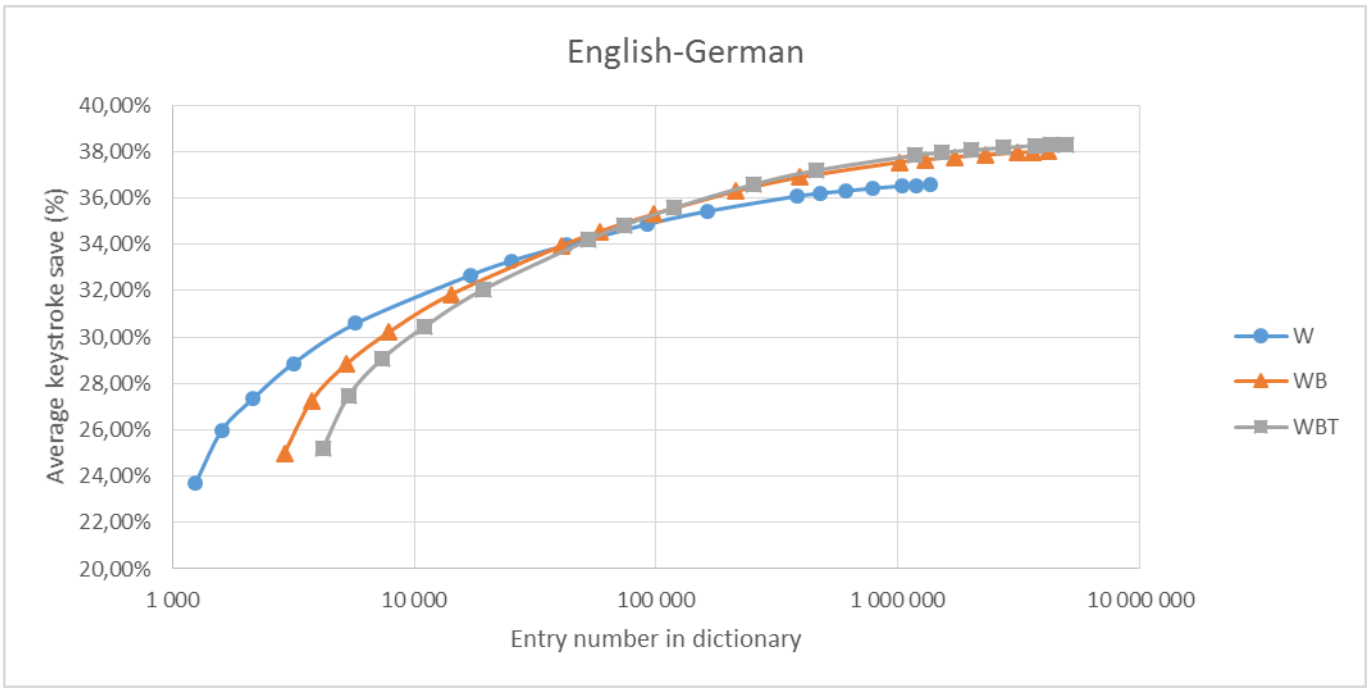

Figure 1. Keystroke saving results for English-German language pair, based on the entry number in the dictionary. For small dictionary sizes the word-only approach is more efficient, but including bigrams and trigrams result in greater keystroke saving beyond 50 000 entries.

legend stands for: W – words; WB – words and bigrams; WBT – words, bigrams and trigrams.

First of all, the curves and their relative positions are similar in the two figures. This means that using physical size of the dictionary gives basically the same results as the number of entries, they are interchangeable.

Secondly, it can be concluded that keystroke saving increases with dictionary size, but its marginal keystroke saving decreases, i.e. it reaches a plateau. The maximum keystroke saving value is about 36.6% for the only-word case, 38.0% for the word-bigram case, and it is around 38.3% for the word-bigram-trigram case for the English-German language pair.

Keystroke saving values are lower for the English-Hungarian language pair, the maximums are around 31-32%. The reason behind this is the fact that the Hungarian language is agglutinative, therefore the number of frequently used text units are higher, making it more difficult to predict the one being typed.

Dictionaries using only words performs better at small sizes (the other curves seems to be shifted right in that interval), but including bigrams and trigrams result in greater keystroke saving after 3 MB or 50k entries. Hence, using only words is better for small, compact dictionaries. On the other hand, state of the art computers can handle a 100 MB dictionary easily, so it's worth including longer text units in the dictionary.

Including trigrams in addition to bigrams changes only a little. This could be caused by the o_{min} threshold being too high, but lowering that threshold causes insufficient memory problems in the pairing and scoring step.

V. Summary

We presented a dictionary based approach for predictive translation typing in this paper. We investigated one aspect of dictionary creation: how does including multi-word text units in the dictionary affect the dictionary size and the keystroke saving. Measurements were carried out on English-German and English-Hungarian language pairs.

It was stated that including bigrams and trigrams along words in a dictionary makes its size bigger, but comes with greater keystroke saving. As state of the art computers can handle even the bigger dictionaries easily, the use of multi-word text units is advisable.

As future study, other kinds of text units could be investigated. For example, the definition could be loosened, and not just consecutive words could be allowed, or the order of the words could be changed.

This dictionary based approach could be compared to monolingual prediction and to the TransType solutions. The sensitivity and optimal interval of the thresholds could also be measured.

References

[1] Barrachina, Sergio, et al. "Statistical approaches to computer-assisted translation." Computational Linguistics 35.1 (2009): 3-28.

[2] Foster, George, Pierre Isabelle, and Pierre Plamondon. "Target-text mediated interactive machine translation." Machine Translation 12.1-2 (1997): 175-194.

[3] Foster, George, Philippe Langlais, and Guy Lapalme. "TransType: text prediction for translators." Proceedings of the second international conference on Human Language Technology Research. Morgan Kaufmann Publishers Inc., 2002.

[4] How, Yijue, and Min-Yen Kan. "Optimizing predictive text entry for short message service on mobile phones." Proceedings of HCII. Vol. 5. 2005.

[5] Masui, Toshiyuki. "An efficient text input method for pen-based computers."Proceedings of the SIGCHI conference on Human factors in computing systems. ACM Press/Addison-Wesley Publishing Co., 1998.

[6] McInnes, Bridget T. Extending the log likelihood measure to improve collocation identification. Diss. UNIVERSITY OF MINNESOTA, 2004.

[7] Silfverberg, Miika, I. Scott MacKenzie, and Panu Korhonen. "Predicting text entry speed on mobile phones." Proceedings of the SIGCHI conference on Human factors in computing systems. ACM, 2000.

[8] Tiedemann, Jörg. "Parallel Data, Tools and Interfaces in OPUS." LREC. 2012.

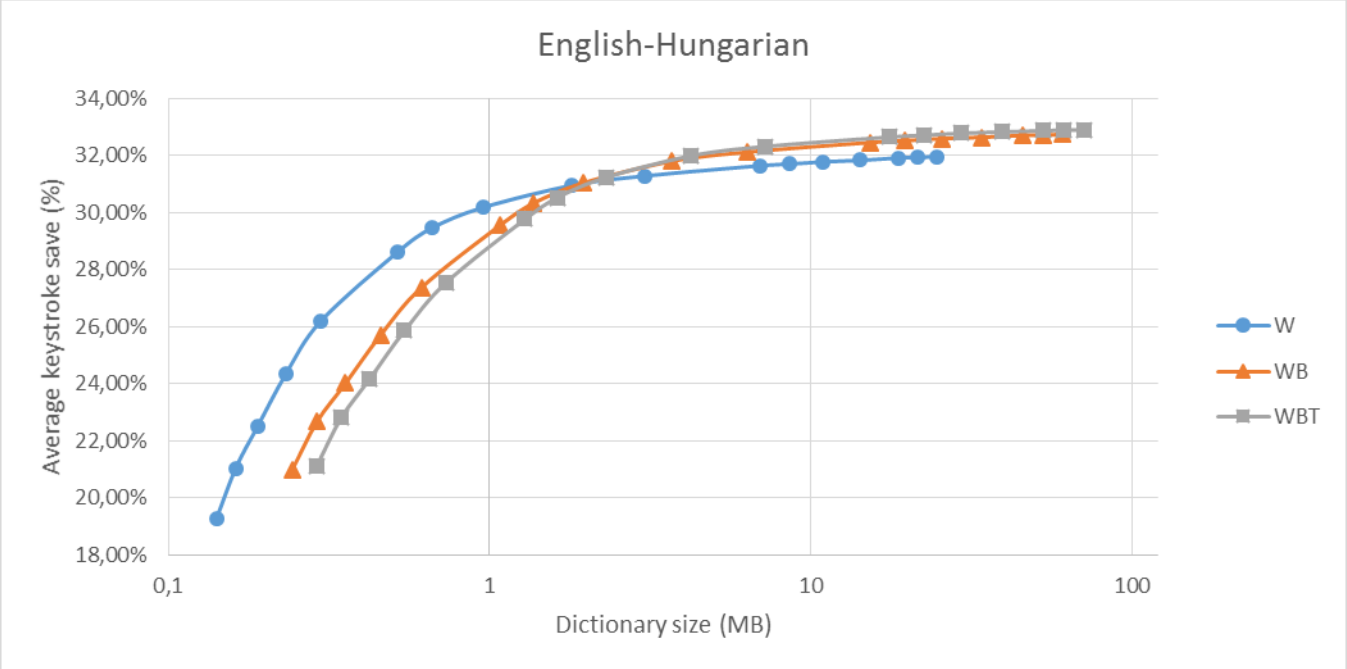

Figure 3. Keystroke saving results for English-Hungarian language pair, based on the physical size of the dictionary. For small dictionary sizes the word-only approach is more efficient, but including bigrams and trigrams result in greater keystroke saving beyond 2 megabytes.