picoTrans: Using Pictures as Input for Machine Translation on Mobile Devices

Andrew Finch
NICT
Multilingual Translation Laboratory
andrew.finch@nict.go.jp

Wei Song
University of Tokyo
Graduate School of Science and Technology
song@cl.ci.i.u-tokyo.ac.jp

Kumiko Tanaka-Ishii
University of Tokyo
Graduate School of Science and Technology
kumiko@i.u-tokyo.ac.jp

Eiichiro Sumita
NICT
Multilingual Translation Laboratory
eiichiro.sumita@nict.go.jp

Abstract

In this paper we present a novel user interface that integrates two popular approaches to language translation for travelers allowing multimodal communication between the parties involved: the picture-book, in which the user simply points to multiple picture icons representing what they want to say, and the statistical machine translation (SMT) system that can translate arbitrary word sequences. Our prototype system tightly couples both processes within a translation framework that inherits many of the the positive features of both approaches, while at the same time mitigating their main weaknesses. Our system differs from traditional approaches in that its mode of input is a sequence of pictures, rather than text or speech. Text in the source language is generated automatically, and is used as a detailed representation of the intended meaning. The picture sequence which not only provides a rapid method to communicate basic concepts but also gives a 'second opinion' on the machine transition output that catches machine translation errors and allows the users to retry the translation, avoiding misunderstandings.

1 Introduction

Handheld mobile devices with touch-screens are becoming increasingly popular, and the flexibility and versatility of the touch-screen gives rise to a panoply of possible user interface designs for linguistic applications. Currently, there is a large demand for applications to aid translation, and this paper proposes a new extension to these applications, implemented in the form of a prototype system *picoTrans* (PICture ICOn TRANSlator) applied to the travel domain.

Our proposal combines machine translation (MT) with the idea of a picture-based translation-aid [Graf, 2009; Meader, 1995; Warrink, 2007; Stillman, 2007; Flanagan, 2008]. Picture-based translation-aids, as will be discussed more in detail in the following section, have been used in paper book forms and are currently integrated into hand-held devices but remain uncombined with MT systems. Briefly, in picoTrans, the user taps picture icons appearing on the touch-screen, just like in a picture-based translation-aid. The system automatically generates the possible sentences from those selected icons, and feeds them to the MT system in order that it can display the translated result. For example, as illustrated in Figure 1, suppose a user wished to translate the expression 'I want to go to the restaurant.', with a the picture book, the user would point at 2 pictures: 'I want to to go to ∼', and 'restaurant'. A similar scenario for our system is shown in Figure 2. Here again the user points to a sequence of icons, however, in our case the sequence of icons is maintained on the display for the users to see, and interact with if necessary. When the input is complete, the system generates the full sentence in the source language automatically, which is then translated by the SMT software and displayed on the screen together with the icon sequence.

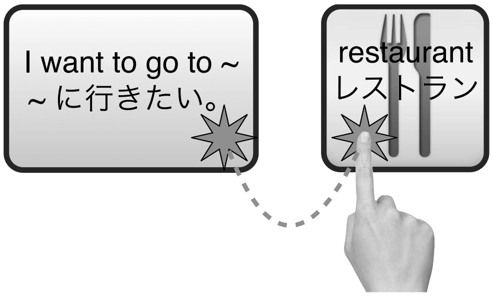

Figure 1: The process of communication by using a picture-based translation aid.

The combination of the two approaches is advantageous from the perspective of both the user interface and machine translation. Firstly, from the user interface viewpoint, the major bottleneck of hand-held devices is the difficulty of text entry [MacKenzie and Tanaka-Ishii, 2007]. There have been many text entry systems proposed for small devices [Sirisena, 2002], but still the entry of full sentences is a cumbersome process. Entry by the tapping of icons allows entry of words in only at most a few taps, decreasing the number of actions needed to enter a sentence, thus increasing the efficiency of entry. Speech input is another input option, but recognition errors are commonplace and can be frustrating to users.

Secondly, from the MT viewpoint the icon sequence representation serves as a means of regularizing the sentences to be translated. This is advantageous for the translation system, since a major cause of translation error arises from use of rare words or infrequent sentence forms. An icon-only approach provides standardized wordings/phrases with no typographical errors and therefore the machine translation will not suffer from errors arising from surface form variation.

The basic premise of our user interface, that sequences of images can convey a meaningful amount of information is directly supported by an interesting study into the effectiveness of using pictures to communicate simple sentences across language barriers [Mihalcea and Leong, 2008]. Using human adequacy scores as a measure, they found that around 76% of the information could be transferred using only a pictorial representation. Furthermore, the Talking Mats project [Murphy and Cameron, 2008], has developed a communication framework consisting of sets of pictures attached to mats to enable people with communication difficulties to communicate. In research into collaborative translation by monolingual users, [Hu *et al.*, 2010] propose an iterative translation scheme where users search for images or weblinks that can be used to annotate sections of text to make its meaning more explicit to another user who does not share the same language. In other related work, [Zhu *et al.*, 2007], demonstrate the usefulness of a text-to-picture transduction process (essentially the converse of our icon-to-text generation process) as a way of automatically expressing the gist of some text in the form of images.

There are various applications available for hand-held devices in terms of either picture-based or MT, but none of them adopt both. In the former area, PictTrans [PicTrans, 2010] only shows picture icons, Yubisashi [Yubisashi, 2010] (meaning *finger-pointing*) plays a spoken audio sound when tapping the icons, but these systems do nothing in terms of language generation which is delegated to the human users. Conversely, there are a substantial number of MT systems proposed for hand-held devices, for example the texTra [TexTra, 2010] text translation system and the voiceTra [VoiceTra, 2010] speech translation system, but as far as we are aware, none of them adopt an icon-driven user input system.

2 Motivation

Traditionally, the phrasebook has been the archetypal translation aid for travelers. However, phrasebooks are essentially a large list of commonly-used expressions, with little or no ability for composition into more complex expressions, and no real capability to express anything that falls outside the gamut of the phrasebook itself. One possible solution for this problem has already been proposed: picture-based translation-aids. These have existed for some time in the form of paper books and have become popular as an alternative to traditional phrasebooks [Graf, 2009; Meader, 1995; Warrink, 2007; Stillman, 2007; Flanagan, 2008; Various, 2005], and MT systems due to the simple, efficient way in which information can be conveyed. Picture-based translation books contain pages of picture icons which represent basic concepts that the user of the book may wish to convey. These picture-based books allow the users to simply point at what that wish to communicate, and as a consequence are very intuitive in use. Using this simple strategy for cross-lingual communication, it is possible to convey simple expressions without the need for sentence-level translation. For example, a book contains many icons similar to those shown in Figure 1, the communicators interact by simply seeing the sequence of pictures together with translations of the words 'go' and 'restaurant'.

While picture-books are simple to use, and intuitive, it is not possible to express all possible phrases that one might wish to use while travelling with a picture book. For example, in the Yubishashi range of books [Various, 2005] there are several different books that are specific to the context in which they should be used (there are books dedicated to: general traveling, food, diving, amour etc.). Furthermore, in most cases the user is effectively confined to choosing from the set of icons that appear on the particular page they are on. That is, they are able to use icons from other pages in theory, but doing so would require a time-consuming search through the book itself to find them. In performing this search, the sequence of icons so far which is only maintained in the memories of the users themselves might be lost. Finally, the icons are designed to fit into translation patterns found on the same page as they occur. They *may* work with other patterns in the book, but due to the nature of natural language it will often make no sense.

Other essential problems lie in the fact that the picture icons in paper book form remain static. Icons that are to be presented on the same page should ideally depend dynamically on the context. Also, the number of icons available in a book is often insufficient, even when limiting the domain only to travel. Our prototype system is currently able to handle around 2000 icons, and being an electronic device has the potential to cope with as many as are required.

These limits in expressive power can be overcome by enhancing such picture-based methods into the form of a dynamic user interface working as an information system. This idea can be studied in isolation by considering the user interface in itself, but since a user interface is an interface of an application having a certain motivation, our idea is to build such an interface for MT, since the objective of a picture-based translation-aid is cross-lingual communication. This also brings advantages to the MT side too, as mentioned in the previous section, by standardizing the vocabulary and phrases. In other words, current picture-based translation-aids curtail the expressiveness of natural language too much, while MT suffers from excessive freedom of expressiveness of input, and the combination has the possibility of finding the balancing point.

3 User Interface

3.1 Overview

The user interaction is made through an interface currently implemented as a prototype working on the Apple iPad mobile tablet, but we believe our interface is applicable to smaller devices with reasonably large touch screens of similar screen resolution such as the iPod touch and iPhone.

A storyboard for the user interface flow is given below:

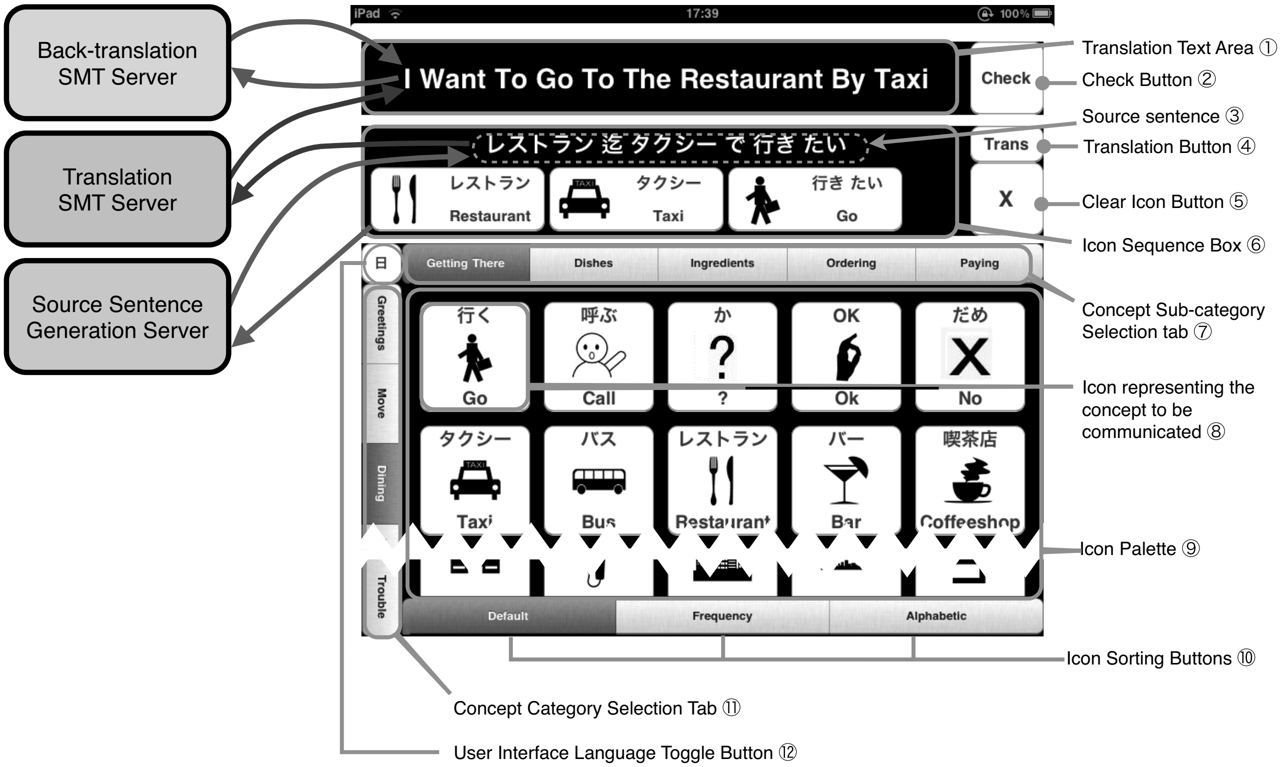

Figure 2: The annotated user interface for the picoTrans system.

(1) The user selects a category for the concept they wish to express

(2) The user selects a sub-category

(3) The user chooses the first icon in the sequence

  a) Go to (1) to select another icon for the sequence

  b) The icon sequence is complete, and the source sentence is acceptable. Continue to step (4)

(4) The user clicks the ‘Trans’ button

(5) The translation appears in the translation text area

3.2 Interface Design

One of the major problems affecting user input on mobile devices is occlusion: the users hands can obscure parts of the user interface and in such cases studies have shown that there can be a considerable decrease in performance [Vogel and Baudisch, 2007; Baur *et al.*, 2010].

With this in mind we have designed the user interface such that the interface elements avoid issues with occlusion. The general layout of the application uses the pocket calculator as a metaphor. The interface groups its interface elements into three distinct *levels*. Each level being physically higher on the device itself as well as performing a higher-level function. The functions at each level correspond directly to the three phrases of user interaction described in the following sections: the selection of icons; the selection and refinement of the source sentence; and translation.

Icon selection functions are performed at the bottom of the device. The area involved with icon sequence display, editing and refinement is immediately above the icon selection area. While interacting with the icon sequence, the user’s hands will obscure parts of the icon selection area, but typically the user will not need to use this area again unless an error has been made. The uppermost interface element is the Translation Text Area. This is never obscured and the user can immediately see their translation, which typically completes in a fraction of a second without needing to move their hand.

Briefly, there are three kinds of interaction required for the user: 1. selection of icons, 2. selection and refinement of the source sentence, and 3. viewing the output translation. The following sections describe each process in more detail.

3.3 Selection of Icons

Icon selection takes place in the icon palette area of the interface (labeled as ⑨ in Figure 2). This contains the set of icons available immediately to the user for translation. The icons are arranged in a rectangular array and can be sorted for efficiency of access.

Icon Categorization

To winnow the choice of icons, the user first selects the icon category in the Icon Category Tab. Let’s say the use selects the category: ‘Dining’. The interface will select the default subcategory from the subcategories for that icon. After the category has been selected the user then either accepts the default sub-category selection, or selects the appropriate sub-category on the sub-category tab. Let us assume the user selects the sub-category: ‘Getting there’. Once the subcategory has been chosen the application displays the icons for that particular sub-category in an order described in the next section. In this example, the user would see icons for ‘Taxi’, ‘Go’ and so on.

Icon Selection and Refinement

The icons on the icon palette can be tapped to add them to the end of the sequence of icons to be translated. The icons have two types of behavior when tapped that depends upon the ambiguity of the underlying content word used to represent them on the interface.

If the underlying content word is unambiguous, the user simply chooses an icon by tapping it. The icon’s changes appearance briefly as feedback to the user that it has been tapped, and a smaller version of the icon appears in the Icon Sequence Box ⑥, on the end of the sequence of icons already present (if any).

If the underlying content word is ambiguous, the user also chooses the icon by tapping it. The icon’s changes appearance briefly as feedback to the user that it has been tapped, and a disambiguation dialog box appears on the screen offering the user several choices for the precise meaning of the icon, this is shown in Figure 3. Often these ambiguous icons are verbs, and the choices are various possibilities for the verb’s form. For example the icon representing the concept ‘go’, might pop up a dialog asking the user to choose between ‘will go’, ‘want to go’, ‘will not go’, ‘went’ etc. Once the user has chosen their preferred semantics for the icon, a smaller version of the icon appears on the end of the sequence of icons already present (if any) in the Icon Sequence Box (⑥ in Figure 2).

The Icon Sequence Box contains the sequence of icons used to express the source sentence to be translated. As the user chooses icons from the icon palette, they appear in sequence from left-to-right in the Icon Sequence Box. An icon sequence is shown in Figure 2.

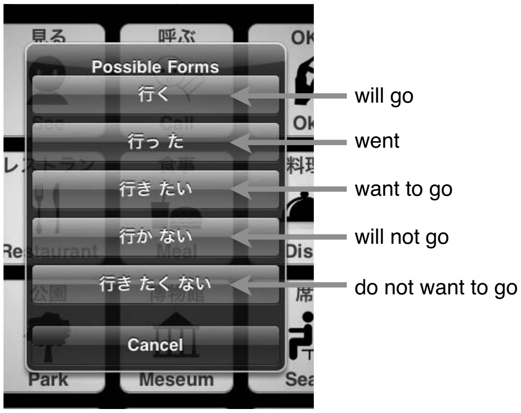

Figure 3: The icon semantics disambiguation dialog.

3.4 Source Sentence Selection and Refinement

Once the icon sequence has been chosen, the user is shown the system’s suggested source sentence for the sequence (③ in Figure 2). The system has been designed so that this source sentence is most likely to be what the user would say.

One aim of the application is to predict the source sentence correctly for the user, however this will not always be possible and the user interface allows the user to interact to refine the semantics of the icon sequence. To do this, the user simply taps on the icon in the Icon Sequence Box (⑥ in Figure 2) that marks the last (rightmost) icon in the sequence of icons that requires modification. The application highlights this icon and pops up a dialog box that allows the user to choose the precise semantics for the icon sequence up to and including the selected icon. The choices presented have been extracted from the source language corpus, and are ordered in terms of language model probability.

3.5 Translation

The user taps the Translation Button (④ in Figure 2) when they are satisfied that the source sentence generated from the (possibly refined) sequence of icons they have selected represents the expression they wish to communicate to the other party. The application then sends the source sentence to the MT server for translation, and the translation appears in the Translation Text Area (① in Figure 2).

There is the risk of misunderstanding, due to the following reasons:

- The MT result is incorrect.
- The user does not understand the translation output in the target language.

In order to tackle mis-understandings arising from MT errors, the user interface has two further user interaction possibilities.

Firstly, the correctness of the translation can be confirmed by back-translation. The user presses the Check Button (② in Figure 2), and the translated text is fed to a MT system that translates from the target language back into the source language. Pressing the Check Button again replaces the back-translation with the translation. The use of back-translation is somewhat controversial since the translation quality can be low and errors may confuse users, but in our system this is mitigated by the high translation quality of our restricted-domain system.

Secondly, the sequence of icons is explicitly displayed on the device for both communicators to see. The picture-based translation-aid in the form of a book has the problems associated with viewing and memorizing the sequences of icons, especially when they are located on different pages. In our system, the icon sequence is explicitly displayed but much more than that, our system is a dynamic information system, and as such offers great potential for enhancing the communication process.

4 System Architecture

Our system consists of 4 components (shown in Figure 2):

- The user interface client
- The source sentence generation server
- The (forward and backward) MT servers.

The user interface client has been described in the previous sections.

Source sentence generation is performed using a language model. The source sentence is generated by concatenating word sequences (phrases) associated with the icons in the system (a phrase is associated with an icon if it contains word/words from a set associated with the icon). The most likely hypothesis according to the language model is chosen as the source sentence to be used. The phrases we use are taken directly from the source language side of the translation model of the SMT system (this table consists of phrases in both source and target languages, together with a set of probabilities for each). Since the MT system uses these phrase-pairs as building blocks in the translation process, by choosing phrases for our icons from this table we are effectively guaranteeing the MT system will be able to translate it.

For our experiments we use CleopATRa [Finch *et al.*, 2007], an in-house SMT decoder that is based on the phrase-based MT techniques introduced by [Koehn *et al.*, 2007], integrating our models within a log-linear framework. Word alignent was performed using GIZA++ [Och and Ney, 2003] and MOSES [Koehn *et al.*, 2007] tools.

A 5-gram language model built with Knesser-Ney smoothing was used. The systems were trained in a standard manner, using a minimum error-rate training (MERT) procedure [Och, 2003].

The SMT systems were trained on approximately 700,000 bi-lingual sentence pairs comprised of the types of expressions typically found in travel phrasebooks t[Kikui *et al.*, 2003]. This is a limited domain, and the sentences in this domain tend to be very short (on average 7-words in the English side of the corpus), making them easy to translate. The MT system is a state-of-the-art system, and as a consequence of limiting the application to short sentences in a restricted domain it is capable of high quality translations; a close relative of this system is being used successfully in commercial MT applications [TexTra, 2010; VoiceTra, 2010].

5 Evaluation

One of the main concerns about the technique proposed in our system is its expressive power within the domain, since sentences need to be expressed by only using icons that are available on the device. We therefore conducted an evaluation of the user interface to determine the proportion of in-domain sentences it was capable of representing. To do this we took a sample of 100 sentences from a set of held-out data drawn from the same sample as the training corpus, and determined whether it was possible to generate a semantically equivalent form of each sentence using the icon-driven interface and its source sentence generation process. The current version of the prototype has not been developed sufficiently to include sets of icons to deal with numerical expressions (prices, phone numbers, dates and times etc.), so we excluded sentences containing numerical expressions from our evaluation set (the evaluation set size was 100 sentences after the exclusion of sentences containing numerical expressions). Handling numerical expressions is relatively straightforward however, and we do not foresee any difficulty in adding this functionality into our system in the future. The set of icons used in the evaluation corresponded to the most 2010 frequent content words in the English side of the training corpus, that is content words that occurred more than 28 times in the corpus. Thus value was chosen such that the number of icons in the user interface was around 2000, a rough estimate of the number of icons necessary to build a useful real-world application. We found that we were able to generate semantically equivalent sentences for 74% of the sentences in our evaluation data, this is shown in Figure 4 together with statistics (based on a 30-sentence random sample from the 100 evaluation sentences) for cases where fewer icons were used. We feel this is an high level of coverage given the simplifications that have been made to the user interface. For 49 of the 74 sentences that we were able to cover with our system (66% of them), the system proposed the correct source sentence to the user, and no icon refinement was necessary.

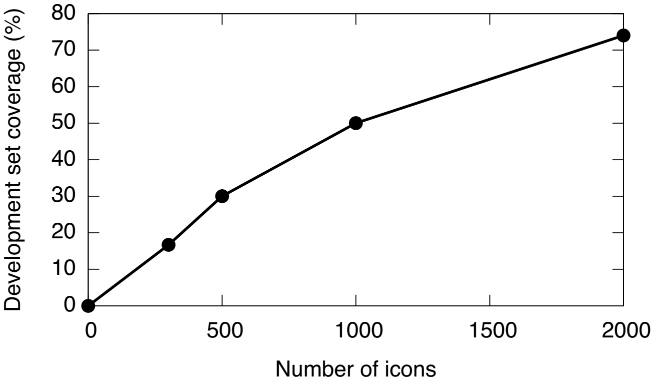

Figure 4: The coverage of unseen data with icon set size.

We also studied the number of key-press actions needed to enter these sentences using icons relative to the number that would have been needed to input them using the device's text-entry interface. Here we assumed that each icon would require 3 key presses to select, but often the icons from the same icon subcategory can be used, and these icons would only require 1 key press. Furthermore, we only count the key presses needed for the romaji input, and not those needed for the kana-kanji conversion step which involves time consuming disambiguation and possible re-entry. So our estimate represents an upper-bound for the number of key press actions necessary for picoTrans, and an underestimate of the number required for direct text entry. However, the time required for one key press isn't equal for icon input and text input, and we did not measure this in our experiments. We also made no attempt to measure of effect of user input errors on the input process. Measuring these factors remains future work. Our measurements include the additional key presses needed to select the semantics of ambiguous icons, and also the key presses necessary to modify the source sentence to have the intended meaning.

In our experiments we found that the icon entry system required only 57% of the number of key press actions of the text entry method: 941 key presses for the icon-driven input method as opposed to 1650 for text entry.

6 Conclusion

In this paper we have presented a novel user interface that integrates ideas from two different paradigms of translation for travelers: picture-books and statistical machine translation. Our approach offers all of the advantages of the simplistic but powerful representation of the picture-books, and at the same time is able to produce natural language in the target language able to unambiguously express the source language user's meaning. The resulting system is both more expressive than the picture-book approach, and at the same time mitigates the problems due to errors in the MT system by facilitating more accurate translation and also providing the users of the system with two independent means of checking the MT accuracy: the icon sequence itself, and the back-translation of the target sentence into the source language. Our evaluation has shown that the icon-based input system covers around 74% of the sentences in the domain of our basic travel expression corpus, and furthermore, can significantly reduce the number of key presses required to enter the expression to be translated relative to a text-only input method.

In future work we would like to explore the possibility of structured input, and also develop a more collaborative environment for the users to interact through the common language that is the icon sequence.

References

[Baur *et al.*, 2010] Dominikus Baur, Sebastian Boring, and Andreas Butz. Rush: repeated recommendations on mobile devices. In *IUI '10: Proceeding of the 14th international conference on Intelligent user interfaces*, pages 91–100, New York, NY, USA, 2010. ACM.

[Finch *et al.*, 2007] Andrew Finch, Etienne Denoual, Hideo Okuma, Michael Paul, Hirofumi Yamamoto, Keiji Yasuda, Ruiqiang Zhang, and Eiichiro Sumita. The nict/atr speech translation system for iwslt 2007. In *In Proceedings of the IWSLT*, Trento, Italy, 2007.

[Flanagan, 2008] Cheryn Flanagan. *Me No Speak: China*. Me No Speak, 2008.

[Graf, 2009] Dieter Graf. *Point it: Traveller's Language Kit - The Original Picture Dictionary - Bigger and Better (English, Spanish, French, Italian, German and Russian Edition)*. Graf Editions, 2009.

[Hu *et al.*, 2010] Chang Hu, Benjamin B. Bederson, and Philip Resnik. Translation by iterative collaboration between monolingual users. In *Proceedings of Graphics Interface 2010*, GI '10, pages 39–46, Toronto, Ont., Canada, Canada, 2010. Canadian Information Processing Society.

[Kikui *et al.*, 2003] G. Kikui, E. Sumita, T. Takezawa, and S. Yamamoto. Creating corpora for speech-to-speech translation. In *Proceedings of EUROSPEECH-03*, pages 381–384, 2003.

[Koehn *et al.*, 2007] Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowa, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondrej Bojar, Alexandra Constantin, and Evan Herbst. Moses: open source toolkit for statistical machine translation. In *ACL 2007: proceedings of demo and poster sessions*, pages 177–180, Prague, Czeck Republic, June 2007.

[MacKenzie and Tanaka-Ishii, 2007] Scott MacKenzie and Kumiko Tanaka-Ishii, editors. *Text Entry Systems — Accessibility, Mobility, Universality—*. Morgan Kaufmann, 2007.

[Meader, 1995] Jonathan Meader. *The Wordless Travel Book: Point at These Pictures to Communicate with Anyone*. Ten Speed Press, 1995.

[Mihalcea and Leong, 2008] Rada Mihalcea and Chee Wee Leong. Toward communicating simple sentences using pictorial representations. *Machine Translation*, 22:153–173, September 2008.

[Murphy and Cameron, 2008] Joan Murphy and Lois Cameron. The effectiveness of talking mats with people with intellectual disability. *British Journal of Learning Disabilities*, 36(4):232–241, 2008.

[Och and Ney, 2003] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[Och, 2003] Franz J. Och. Minimum error rate training for statistical machine trainslation. In *Proceedings of the ACL*, 2003.

[PicTrans, 2010] PicTrans. *PicTrans - a simple picture-based translation system*. 7Zillion, 2010. http://www.7zillion.com/iPhone/PicTrans/.

[Sirisena, 2002] Amal Sirisena. Mobile text entry. *Retrieved July*, 27, 2002.

[Stillman, 2007] Stillman. *Kwikpoint International Translator (English, Spanish, French, Italian, German, Japanese, Russian, Ukrainian, Chinese, Hindi, Tamil, Telug, Kannada, Malayalam, Gujarati, Bengali and Korean Edition)*. Kwikpoint, 2007.

[TexTra, 2010] TexTra. *TexTra (Text Translator by NICT)*. NICT, 2010. http://mastar.jp/translation/textra-en.html.

[Various, 2005] Various. *Traveling Pointing Book*. Information Center Publishing, 2005.

[Vogel and Baudisch, 2007] Daniel Vogel and Patrick Baudisch. Shift: a technique for operating pen-based interfaces using touch. In *PROC. CHI '07*, pages 657–666. ACM Press, 2007.

[VoiceTra, 2010] VoiceTra. *VoiceTra (Voice Translator by NICT)*. NICT, 2010. http://mastar.jp/translation/voicetra-en.html.

[Warrink, 2007] Gosia Warrink. *ICOON Global Picture Dictionary (English, Spanish, French, Italian, German, Japanese, Russian, Chinese and Hindi Edition)*. Amberpress, 2007.

[Yubisashi, 2010] Yubisashi. *Yubisashi*. Information Center Publishing, 2010. Available in many languages, found at `http://www.yubisashi.com/free/t/iphone/`, visited in 2010, August.

[Zhu *et al.*, 2007] Xiaojin Zhu, Andrew B. Goldberg, Mohamed Eldawy, Charles R. Dyer, and Bradley Strock. A text-to-picture synthesis system for augmenting communication. *In Proceedings of the 22nd International conference on Artificial intelligence*, 2:1590–1595, 2007.