Source Language Generation from Pictures for Machine Translation on Mobile Devices

Andrew Finch[1], Wei Song[2], Kumiko Tanana-Ishii[2], and Eiichiro Sumita[1]

[1] NICT, Keihanna Science City, Kyoto, 619-0289, Japan.
{andrew.finch,eiichiro.sumita}@nict.go.jp
http://kccc.nict.go.jp
[2] University of Tokyo, Tokyo, 101-0021, Japan.
{song@cl.ci.,kumiko@}i.u-tokyo.ac.jp
http://www.cl.ci.i.u-tokyo.ac.jp

Abstract. This paper proposes a new method for the generation of natural language from a sequence of pictures and shows its application to machine translation within the framework of the picoTrans system. The picoTrans system is an icon-driven user interface for machine translation that facilitates cross-lingual communication through two heterogeneous channels of communication simultaneously. The first channel being the usual automatic natural language translation method; the second channel being a sequence of pictures that both parties understand which conveys structured semantic information in parallel with the first channel. Users are able to communicate using this device both by using it as a picture book and also by using it as a machine translation device. By pointing at pictures alone, basic expressions can often be communicated, eliminating the need for machine translation altogether, and even with machine translation, the picture sequence provides a useful second opinion on the translation that helps to mitigate machine translation errors. There are limits, however to the expressiveness of a sequence of pictures compared to the expressiveness of natural language. This paper looks at two methods by which syntactic information can be added into a sequence of pictures: a hidden n-gram model and monotonic transduction using a phrase-based statistical machine translation system. This additional information is added automatically, but the system allows the user to interact to refine the generated language. We evaluate both methods on the task of source sentence generation in Japanese using automatic machine translation evaluation metrics, and find the statistical machine translation method to be the more effective technique.

Keywords: user interface, machine translation, mobile devices

1 Introduction

Recently there has been a huge increase in the demand for machine translation (MT) services, as the translation quality for many language pairs has improved to levels that are of practical use. One common platform for applications of machine translation is mobile devices, since they can be used wherever they are

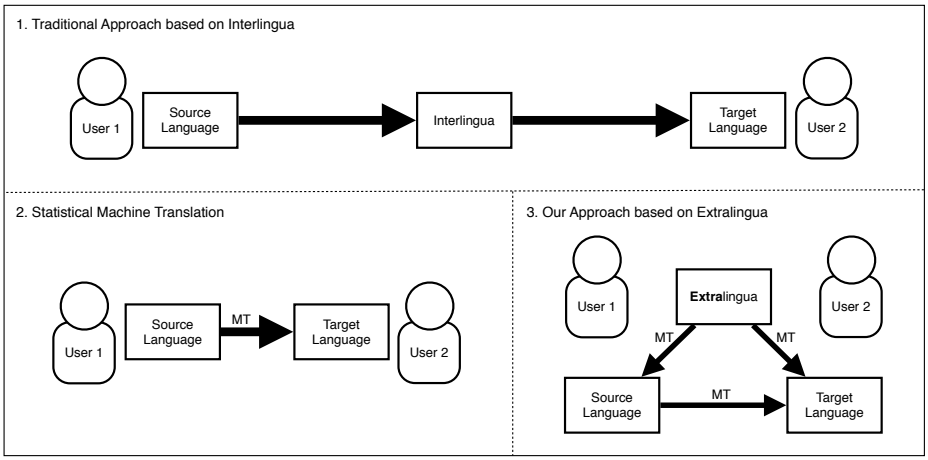

Fig. 1. Various Translation Channels Among Communicators Using Machine Translation

needed. The popularity of MT systems has cast light on the marginal problems of MT other than its translation quality. A major issue facing real world MT applications is the sheer variety of input that user could feed into an MT system. For example the input may contain mistakes, irregular or ungrammatical sentences, abbreviated words, smileys and contractions of words (such as writing "u" instead of "you"). These phenomena will degrade the performance of an MT system, but are essentially issues external to the core problem the MT system has to tackle, which is already a hard problem.

One way to address these issues would be to collect a large corpus consisting of irregular usages together with the corresponding correct usages, and learn to regularize the language in a supervised fashion. However, this approach is obviously limited, since schemes through which real users irregularize can depend on the user's communication mode or code, context, and even on his/her wit.

The picoTrans system [1] offers a simple solution: adopting an *extralingua* to assist MT. We have chosen the term *extralingua* deliberately due to its relationship to the term *interlingua*, and we expand on this later in the paper. Figure 1, shows some methods proposed for accomplishing translation. In part (1) of the figure, translation is performed through an interlingua, an intermediate language placed in between the source and the target language. When the interlingua is a natural language, the communication channel can be a concatenation of two MT systems: the first from source to interlingua, the second from interlingua to target [2]. Part (2) of the figure shows a process of direct translation from source to target. This can be achieved using widely-studied, state-of-the-art statistical machine translation systems.

In contrast, our approach (3) in Figure 1 uses an *extralingua*, which is exposed to both communicators. Both users are able to interact with the extralingua, assisted by three MT systems: the first between the extralingua to the source

language, the second between the source language and the target language, and the third between the extralingua and the target language. The reader might wonder why MT is needed at all if such an extralingua exists. This is in fact the point: the communicators lack in a common language through which they can communicate, and so far we have only considered ways to bridge this gap by using just a single MT channel. However, under many circumstances, the communicators do have other means for communication, such as images, signs and actions and will often use them when other means fail. This other mode of communication can be adopted in parallel independently of the MT channel, but our idea is to investigate the tight coupling of a second communication channel directly into a machine translation system.

The basic premise of our user interface, that sequences of images can convey a meaningful amount of information is directly supported by the findings of a number of studies. In [3], the effectiveness of using pictures to communicate simple sentences across language barriers is assessed. Using human adequacy scores as a measure, they found that around 76% of the information could be transferred using only a pictorial representation. In language generation [4] explore the possibility of communication by means of concepts. In assistive communication, the Talking Mats project [5], has developed a communication framework consisting of sets of pictures attached to mats to enable people with communication difficulties to communicate. There are other related systems and ideas based on the principle of using pictorial communication as a linguistic aid [6, 7]. In [8], a method for transforming icons into speech is proposed to aid people with Cerebral Palsy in communication.

In research into collaborative translation by monolingual users, [9] 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, [10], 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.

In our approach, we adopt icon sequences as the primary mode of input. There are multiple advantages to doing this. First and above all is to improve the quality of communication between users. Adopting an extralingua allows the users to communicate via two heterogeneous channels. Since we cannot expect MT output to be perfect, having a second independent mode of communication to reinforce or contradict will lead to a greater mutual understanding. Secondly, we believe that by constraining the user's input it should be possible to improve the MT quality since the input becomes regularized as a consequence, reducing variance in the input sequence and also decreasing the number of unexpected entries. This idea is supported by a study showing that normalizing language using a paraphraser can lead to improvements in translation quality [11].

A simple example of using a picture-book to communicate is illustrated in Figure 2. Suppose a user wished to translate the expression 'I want to go to the restaurant'; with a picture book, the user might point at 2 pictures: 'I want to to go to ∼', and 'restaurant'. A similar scenario for the picoTrans system is shown in Figure 3.

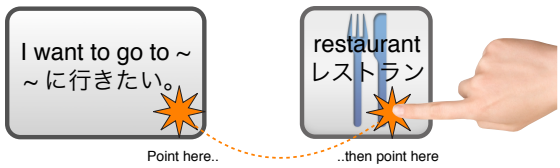

Fig. 2. The process of communication by using a picture-based translation aid for an English person. The order in which the pictures are pointed at is important, and thus if a Japanese person wished to convey the same information, the pointing order would be reversed (a Japanese user would first point to the 'restaurant' icon and then to the 'I want to go to' icon). This directly reflects differences in word order between the languages.

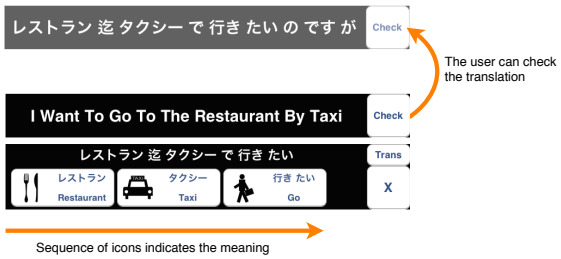

Fig. 3. The process of communication by using the picoTrans system. The sentence is a little more complex than that in Figure 2 to illustrate the additional expressive power of our approach. The picoTrans system displays the icon sequence, together with a translation of the user's intended meaning. This translation can be checked in the target language by referring to the icon sequence, and in the source language by referring to a back-translation.

In the next Section we describe our prototype system picoTrans; for a more complete description of the interface and operation of this system the reader is referred to [1]. The following section addresses the issue of natural language generation from an icon sequence, and describes the two approaches we have studied. Then we present an extension of the experiments reported in [1] that measure the expressiveness of the icon-based input approach on a limited domain, and also present our evaluation of the source generation techniques. Finally we conclude and offer some avenues for further research.

2 The picoTrans System

Picture-based translation-aids have been used in paper book forms and are currently integrated into hand-held devices but remain uncombined with machine translation systems. Briefly, in our proposed system 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 machine translation in order that it

can display the translated result. Unlike a picture book, 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 machine translation software and displayed on the screen together with the icon sequence. The user interaction is made through an interface which is currently implemented as a prototype working on the Apple iPad mobile tablet, although we believe our interface is applicable to smaller devices with touch screens such as mobile phones.

When communicating through our user interface, the user may combine the pictures in considerably more combinations than is possible with a picture book designed with combinations from only within the same page spread of the book in mind, making the application more expressive than a book. The machine translation system can contribute a detailed and precise translation which is supported by the picture-based mode 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 sentence, avoiding misunderstandings.

Note how such a system is advantageous when applied to MT on mobile devices; the user input on these mobile devices can be cumbersome in the case of textual input [12], or errorful in the case of speech input [13]. As a consequence some users prefer to use simpler, more dependable methods of cross-lingual communication such as the picture book translation aids which are becoming increasingly popular both in paper form, and also in the form of electronic translation aid applications.

There are various applications available for hand-held devices in terms of either picture-based or machine translation, but none of them adopt both. In the former area, PictTrans [14] only shows picture icons, Yubisashi [15] (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 [16] text translation system and the voiceTra [17] speech translation system, but as far as we are aware, none of them adopt an icon-driven user input system.

2.1 User Interface

A diagram of the full user interface for the picoTrans prototype system is shown in Figure 4. In brief, we allow the user to input what they wish to express as a sequence of bi-lingually annotated icons. This is in essence the same idea as the picture-book. Users can switch the user interface into their own language by pressing the User Interface Language Toggle Button (⑫ in Figure 4). The translation process proceeds as follows:

(1) The user selects a category for the concept they wish to express ⑪
(2) The user selects a sub-category ⑦
(3) The user chooses an icon ⑨, which is appended to the icon sequence ⑥

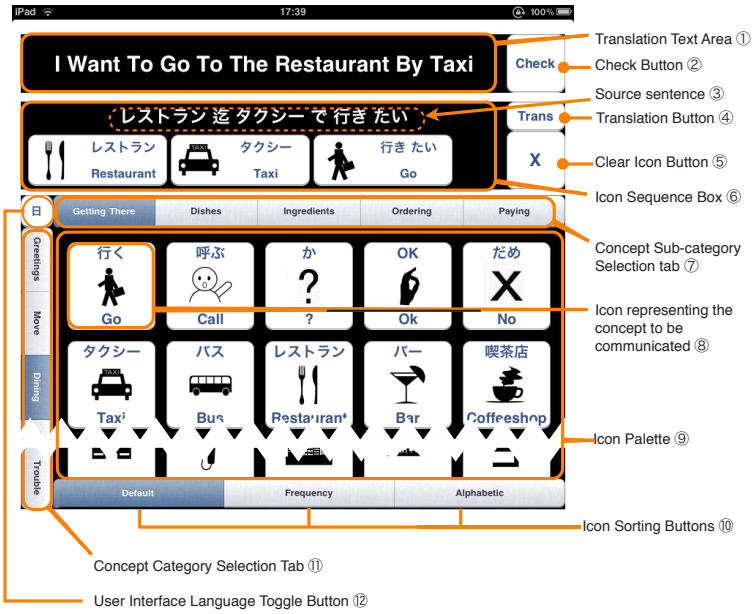

Fig. 4. The annotated user interface for the picoTrans system.

a) Go to (1) to select another icon for the sequence
b) The icon sequence is complete. Continue to step (4)

(4) The user interacts with the system to refine the source sentence (described in Section 3.4)
(5) The user clicks the 'Trans' button ④
(6) The translation appears in the translation text area ①

Once translation has completed, pressing the Check Button ② shows the back-translation of the target sentence into the source language for the user to verify the translation. Pressing the 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.

3 Source Sentence Generation

3.1 Language model approach

In previous work, conducted on Japanese input [1], the icon sequence was transformed into natural language by using a simple language-model-based approach to restore the missing function words. This approach was well-suited to the Japanese language which is quite regular in form, uses particles adjacent to content words to indicate their function, and contains no determiners. Bi-grams

containing pairs of content words with function words attached to either left or right were extracted from training data, and these bigrams were inserted in place of their corresponding content words in the generation process. The model proposed in [1] used a 5-gram language model to score the set of hypotheses resulting from all possible such substitutions, and selected the hypotheses with highest language model score as the best candidate. A beam search was employed to reduce the search space to a manageable size. While this approach proved very effective for Japanese where the simple addition of particles can, for many simple sentences, produce a good Japanese sentence from a sequence of content words representing the icons being input, we believe in order to generate other less suitable natural languages from sequences of icons, a more general approach is necessary. For this we turn to statistical machine translation.

3.2 Machine Translation

The task of transforming our icon sequence into the full source sentence is quite similar to the task of transliteration generation which can be performed using a phrase-based statistical machine translation system (SMT) using a monotonic constraint on the word re-ordering process [18, 19]. We adopt a similar approach, but use a Bayesian co-segmentation technique (explained in the next section) to derive the phrase table for the SMT system.

In order to train our SMT system, we generate a training corpus by means of word deletion. In our experiments we used Japanese as the source language and we analyzed our corpus using the publicly available MeCab [20] morphological analysis tool. A set of POS tags representing the classes of content words that would be represented by icons in our system (for example nouns, verbs, adjectives etc.) was compiled by hand, and the remaining classes of words (particles, dependent nouns and auxiliary verbs) were deleted from the source sentence. Furthermore, all inflected lexemes were reduced to their lemmata. The result of this process was a bilingual corpus, consisting of a sequence of content words (in lemma form) on the source side representing the icon sequence, and the full source sentence on the target side (see Figure 5).

For our experiments we used a phrase-based machine translation decoder closely related to the MOSES [21] decoder, integrating our models within a log-linear framework [22]. Phrase-pair discovery and extraction was performed using a Bayesian bilingual aligner [23]. 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 [24] with respect to the BLEU score [25] on the held-out development data to optimize the log-linear model weights.

The machine translation systems were trained on approximately 700,000 bilingual sentence pairs comprised of the types of expressions typically found in travel phrase books. This is a very 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 very easy to translate. The machine translation 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 translation.

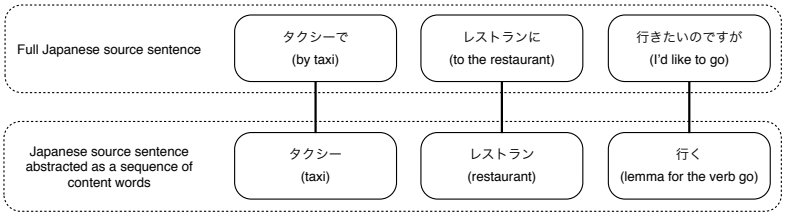

Fig. 5. Co-segmentation of a sequence of content words representing icons in our system, with the corresponding full Japanese sentence that they represent.

3.3 Bayesian Co-segmentation

At the core of all phrase-based statistical machine translation systems is the phrase-table. This table is the basic set of building blocks that are used to construct the translation.

The creation of a phrase-table during a typical training procedure for a phrase-based SMT system consists of word alignment step, often using GIZA++ [26], followed by a phrase-pair extraction step using heuristics (for example *grow-diag-final-and* from the MOSES [21] toolkit). This approach works very well in practice, but is asymmetrical with respect to source and target and is based on maximum likelihood methods that tend to over-fit the data.

The model we use for co-segmentation is based on a Dirichlet process model, similar to approach of [23]. We use a Bayesian approach here not only because results show that this approach is more effective on monotonically alignable sequences than using GIZA++/MOSES heuristics [23], but also because it results in a single self-consistent bilingual segmentation of the corpus. We believe this consistency is a very desirable characteristic for building our models since our system generates natural language simply by composing these phrase-pairs. This co-segmentation process is illustrated in Figure 5.

3.4 User Interaction

The output from both methods of source language generation are search graphs that represent the process by which the source sentence was constructed. This graph is provided by the MT system or n-gram model to the user interface client, which is able to use the information in the graph to guide the user to a satisfactory outcome without the need for continuous re-decoding of the input during the interaction process. This interaction process has not been developed very far in this work and remains an interesting area for future study. In related work [27], it has been demonstrated that an MT search graph can be used effectively in an interactive manner to assist human translation.

In our system, following the generation process from the icon sequence, the user is presented with the most probable hypothesis for the full source sentence given the input sequence. Should this sentence not convey the user's intended meaning, the user is able to interact with the icon sequence in order to refine the generated sentence. The user may tap on any icon in the sequence of icons

displayed on the interface. The user interface will consult the search graph and present the user with an n-best list of partial translation hypotheses up to and including the translation of the icon that was selected. At present we do not allow direct text entry into the system, although we do appreciate this would be possible and perhaps necessary in a real-world system, as our research is primarily concerned with exploring the possibilities arising from an icon-based approach to user input. The price to be paid by restricting the input in this manner is expressiveness, and we therefore examine our system empirically with this in mind in the following section, which is an extension of the experiments reported in [1].

4 Evaluation

4.1 Expressive Power

One of our main concerns about icon-driven user input was 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 system 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 them 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 6 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 a high level of coverage given the simplifications that have been made to the user interface. A comparison of the two methods without human correction of the output are shown in Figure 7. We found that the MT method gave a higher level of coverage.

4.2 Quality of Source Sentence Generation

We measured the quality of the source language generation component of our system using version 13a of the NIST mteval scoring script in terms of the BLEU

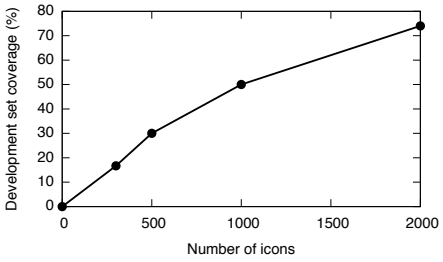

Fig. 6. The coverage of unseen data with icon set size, with human interaction.

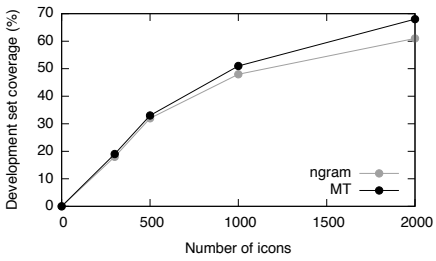

Fig. 7. The coverage of unseen data with icon set size, without human interaction.

score [25], and the NIST score [28], both common methods for measuring machine translation performance based on n-gram precision. We filtered sentences containing numerical expressions from our evaluation data; the initial set of 510 Japanese sentence was reduced to 455 sentences after filtering. We fed these sentences through the MeCab morphological analyzer and removed the words that will not be associated with icons in our systems. Our experimental results are shown in Table 1. The scores for the hidden n-gram (lemmatized) and SMT generation were derived from exactly the same input. The score of the hidden n-gram system is low in this case because it is unable to generate the inflected forms. Therefore, in a second experiment we allow the hidden n-gram model to generate from the correct surface forms of the inflected words. This gives the model an unfair advantage over the SMT generation model which needs to predict the inflection. Nonetheless, the source sentences generated by the SMT process score higher in terms of both of the evaluation metrics used. This combined with the fact that we expect it to be generally applicable to all languages, makes it unreservedly the better generation technique for our purposes.

Model	BLEU	NIST
Hidden n-gram (lemmatized)	0.27	4.16
Hidden n-gram (surface form)	0.66	8.60
SMT Generation	0.76	8.89

Table 1. Source sentence generation quality using an SMT approach relative to a hidden-ngram method to restore missing particles.

5 Conclusion

In this paper we have proposed a method of natural language generation from icon sequences based on a phase-based statistical machine translation system coupled with a Bayesian co-segmentation scheme that is used to perform icon-sequence to word-sequence alignment in order to train the translation model. We

have evaluated this method against an n-gram particle-insertion model used previously to generate Japanese natural language from icon sequences. Our results show that the SMT system is able to outperform the n-gram model in terms of both its coverage of the language and also in terms of the quality of the language generated measured using automatic machine translation metrics.

The picoTrans prototype opens up a wide range of possible directions for future research. In future work we plan to investigate the use of more informative abstract representations and the effects of tightly coupling them into the translation process, and also advance our system from a user interface perspective. We would also like to look at the possible ways in which users of the system might interact, with each other bilingually through the system, and also with the system itself monolingually. We plan to enhance the icon selection component of the interface with a more sophisticated algorithm able to dynamically predict an optimal ordering of the icons presented to the user based on their likelihood of being the next choice given the current context of the dialog, geographical location, and the user's history of icon choices. Finally, we believe the icon sequence to natural language generation techniques we are developing may find application outside the domain of machine translation, and this is something we wish to explore in the future, for example in the field of assistive communication aids.

References

1. Song, W., Finch, A.M., Tanaka-Ishii, K., Sumita, E.: picotrans: an icon-driven user interface for machine translation on mobile devices. In: Proceedings of the 16th international conference on Intelligent user interfaces. IUI '11, New York, NY, USA, ACM (2011) 23–32
2. Paul, M., Yamamoto, H., Sumita, E., Nakamura, S.: On the importance of pivot language selection for statistical machine translation. In: HLT-NAACL (Short Papers). (2009) 221–224
3. Mihalcea, R., Leong, C.W.: Toward communicating simple sentences using pictorial representations. Machine Translation **22** (2008) 153–173
4. Zock, M., Sabatier, P., Jakubiec, L.: Message composition based on concepts and goals. International Journal of Speech Technology **11** (2008) 181–193
5. Murphy, J., Cameron, L.: The effectiveness of talking mats with people with intellectual disability. British Journal of Learning Disabilities **36** (2008) 232–241
6. Ader, M., Blache, P., Rauzy, S.: Overcoming communication difficulties: a platform for alternative communication. 4mes Journes Internationales 'L'Interface des Mondes Rels et Virtuels (2008)
7. Power, R., Power, R., Scott, D., Scott, D., Evans, R., Evans, R.: What you see is what you meant: direct knowledge editing with natural language feedback (1998)
8. Vaillant, P., Checler, M.: Intelligent voice prosthesis: Converting icons into natural language sentences. Computing Research Repository **abs/cmp-lg** (1995)
9. Hu, C., Bederson, B.B., Resnik, P.: Translation by iterative collaboration between monolingual users. In: Proceedings of Graphics Interface 2010. GI '10, Toronto, Ont., Canada, Canada, Canadian Information Processing Society (2010) 39–46
10. Zhu, X., Goldberg, A.B., Eldawy, M., Dyer, C.R., Strock, B.: A text-to-picture synthesis system for augmenting communication. In Proceedings of the 22nd International conference on Artificial intelligence **2** (2007) 1590–1595

11. Watanabe, T., Shimohata, M., Sumita, E.: Statistical machine translation on paraphrased corpora. In: Third International Conference on Language Resources and Evaluation (LREC 2002), Las Palmas, Canary Islands (2002)
12. MacKenzie, S., Tanaka-Ishii, K., eds.: Text Entry Systems — Accessibility, Mobility, Universality. Morgan Kaufmann (2007)
13. Suhm, B.: Empirical evaluation of interactive multimodal error correction. In: in IEEE Workshop on Speech recognition and understanding, IEEE, IEEE (1997) 583–590
14. picTrans: A simple picture-based translation system. 7Zillion (2010) http://www.7zillion.com/iPhone/PicTrans/.
15. Yubisashi: Yubisashi. Information Center Publishing (2010) Available in many languages, found at `http://www.yubisashi.com/free/t/iphone/`, visited in 2010, August.
16. TexTra: (Text Translator by NICT). NICT (2010) http://mastar.jp/translation/textra-en.html.
17. VoiceTra: (Voice Translator by NICT). NICT (2010) http://mastar.jp/translation/voicetra-en.html.
18. Finch, A., Sumita, E.: Phrase-based machine transliteration. In: Proc. 3rd International Joint Conference on NLP. Volume 1., Hyderabad, India (2008)
19. Rama, T., Gali, K.: Modeling machine transliteration as a phrase based statistical machine translation problem. In: NEWS '09: Proceedings of the 2009 Named Entities Workshop: Shared Task on Transliteration, Morristown, NJ, USA, Association for Computational Linguistics (2009) 124–127
20. Kudo, T.: MeCab. [Online] Available: http://mecab.sourceforge.net/ (2008)
21. Koehn, P., Hoang, H., Birch, A., Callison-Burch, C., Federico, M., Bertoldi, N., Cowa, B., Shen, W., Moran, C., Zens, R., Dyer, C., Bojar, O., Constantin, A., Herbst, E.: Moses: open source toolkit for statistical machine translation. In: ACL 2007: proceedings of demo and poster sessions, Prague, Czeck Republic (2007) 177–180
22. Och, F.J., Ney, H.: Discriminative training and maximum entropy models for statistical machine translation. In: In Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics (ACL 2002). (2002) 295–302
23. Finch, A., Sumita, E.: A bayesian model of bilingual segmentation for transliteration. In: In Proceedings of the IWSLT, Paris, France (2010)
24. Och, F.J.: Minimum error rate training for statistical machine translation. In: Proceedings of the ACL. (2003)
25. Papineni, K., Roukos, S., Ward, T., Zhu, W.J.: Bleu: a method for automatic evaluation of machine translation. In: ACL '02: Proceedings of the 40th Annual Meeting on Association for Computational Linguistics, Morristown, NJ, USA, Association for Computational Linguistics (2001) 311–318
26. Och, F.J., Ney, H.: A systematic comparison of various statistical alignment models. Computational Linguistics **29** (2003) 19–51
27. Koehn, P.: A web-based interactive computer aided translation tool. In: Proceedings of the ACL-IJCNLP 2009 Software Demonstrations, Suntec, Singapore, Association for Computational Linguistics (2009) 17–20
28. Doddington, G.: Automatic Evaluation of Machine Translation Quality Using N-gram Co-Occurrence Statistics. In: Proceedings of the HLT Conference, San Diego, California (2002)