Automatic Czech – Sign Speech Translation *

Jakub Kanis[1] and Luděk Müller[1]

Univ. of West Bohemia, Faculty of Applied Sciences, Dept. of Cybernetics
Univerzitní 8, 306 14 Pilsen, Czech Republic
{jkanis,muller}@kky.zcu.cz

Abstract. This paper is devoted to the problem of automatic translation between Czech and SC in both directions. We introduced our simple monotone phrase-based decoder - **SiMPaD** suitable for fast translation and compared its results with the results of the state-of-the-art phrase-based decoder - **MOSES**. We compare the translation accuracy of hand-crafted and automatically derived phrases and introduce a "class-based" language model and post-processing step in order to increase the translation accuracy according to several criteria. Finally, we use the described methods and decoding techniques in the task of SC to Czech automatic translation and report the first results for this direction.

1 Introduction

In the scope of this paper, we are using the term Sign Speech (SS) for both the Czech Sign Language (CSE) and Signed Czech (SC). The CSE is a natural and adequate communication form and a primary communication tool of the hearing-impaired people in the Czech Republic. It is composed of the specific visual-spatial resources, i.e. hand shapes (manual signals), movements, facial expressions, head and upper part of the body positions (non-manual signals). It is not derived from or based on any spoken language. CSE has basic language attributes, i.e. system of signs, double articulation, peculiarity and historical dimension, and has its own lexical and grammatical structure. On the other hand the SC was introduced as an artificial language system derived from the spoken Czech language to facilitate communication between deaf and hearing people. SC uses grammatical and lexical resources of the Czech language. During the SC production, the Czech sentence is audibly or inaudibly articulated and simultaneously with the articulation the CSE signs of all individual words of the sentence are signed.

The using of written language instead of spoken one is a wrong idea in the case of the Deaf. Hence, the Deaf have problems with the majority language understanding when they are reading a written text. The majority language is the second language of the Deaf and its use by the deaf community is only particular. Thus the majority language translation to the sign speech is highly important

* Support for this work was provided by the GA of the ASCR and the MEYS of the Czech Republic under projects No. 1ET101470416 and MŠMT LC536.

for better Deaf orientation in the majority language speaking world. Currently human interpreters provide this translation, but their service is expensive and not always available. A full dialog system (with ASR and Text-to-Sign-Speech (TTSS) [1] systems on one side (from spoken to sign language) and Automatic-Sign-Speech-Recognition (ASSR) and TTS systems on second side (from sign to spoken language)) represents a solution which does not intent to fully replace the interpreters, but its aim is to help in everyday communication in selected constraint domains such as post office, health care, traveling, etc. An important part of TTSS (conversion of written text to SS utterance (animation of avatar)) and ASSR systems is an automatic translation system which is able to make an automatic translation between the majority and the sign language.

In rest of this paper we describe our phrase-based translation system (for both directions: Czech to SC and SC to Czech). We compare the translation accuracy of a translation system based on phrases manually defined in the process of training corpus creation (phrases defined by annotators) with accuracy of the system based on phrases automatically derived from the corpus in the training process of Moses decoder [2]. In addition, we introduce a "class-based" language model based on the semantic annotation of the corpus and the post-processing method for Czech to SC translation.

2 Phrase-Based Machine Translation

The machine translation model is based on the noisy channel model scheme. When we apply the Bayes rule on the translation probability $p(\mathbf{t}|\mathbf{s})$ for translating a sentence $\mathbf{s}$ in a source language into a sentence $\mathbf{t}$ in a target language we obtain:

$$\text{argmax}_t p(\mathbf{t}|\mathbf{s}) = \text{argmax}_t p(\mathbf{s}|\mathbf{t})p(\mathbf{t})$$

Thus the translation probability $p(\mathbf{t}|\mathbf{s})$ is decomposed into two separate models: a translation model $p(\mathbf{s}|\mathbf{t})$ and a language model $p(\mathbf{t})$ that can be modeled independently. In the case of phrase-based translation the source sentence $\mathbf{s}$ is segmented into a sequence of I phrases $\bar{\mathbf{s}}_1^I$ (all possible segmentations has the same probability). Each source phrase $\bar{s}_i, i = 1, 2, ..., I$ is translated into a target phrase $\bar{t}_i$ in the decoding process. This particular ith translation is modeled by a probability distribution $\phi(\bar{s}_i|\bar{t}_i)$. The target phrases can be reordered to get more precise translation. The reordering of the target phrases can be modeled by a relative distortion probability distribution $d(a_i - b_{i-1})$ as in [3], where a_i denotes the start position of the source phrase which was translated into the ith target phrase, and b_{i-1} denotes the end position of the source phrase translated into the $(i-1)th$ target phrase. Also a simpler distortion model $d(a_i - b_{i-1}) = \alpha^{|a_i - b_{i-1} - 1|}$ [3], where α is a predefined constant, can be taken. The best target output sentence $\mathbf{t_{best}}$ for a given source sentence $\mathbf{s}$ then can be acquired as:

$$\mathbf{t_{best}} = \text{argmax}_t p(\mathbf{t}|\mathbf{s}) = \prod_{i=1}^{I} [\phi(\bar{s}_i|\bar{t}_i) d(a_i - b_{i-1})] p_{LM}(\mathbf{t})$$

Where $p_{LM}(\mathbf{t})$ is a language model of the target language (usually a trigram model with some smoothing usually built from a huge portion of target language texts). Note, that more sophisticated model in [3] uses more probabilities as will be given in Section 4.1

3 Tools and Evaluation Methodology

3.1 Data

The main resource for the statistical machine translation is a parallel corpus which contains parallel texts of both the source and the target language. Acquisition of such a corpus in case of SS is complicated by the absence of the official written form of both the CSE and the SC. Therefore for all our experiments we use the Czech to Signed Czech (CSC) parallel corpus ([4]).

The CSC corpus contains 1130 dialogs from telephone communication between customer and operator in a train timetable information center. The parallel corpus was created by semantic annotation of several hundreds of dialog and by adding the SC translation of all dialogs. A SC sentence is written as a sequence of CSE signs. The whole CSC corpus contains 16 066 parallel sentences, 110 033 running words and 109 572 running signs, 4082 unique words and 720 unique signs. Every sentence of the CSC corpus has assigned the written form of the SC translation, a type of the dialog act, and its semantic meaning in a form of semantic annotation. For example (we use English literacy translation) for Czech sentence: *good day I want to know how me it is going in Saturday morning to brno* we have the SC translation: *good_day I want know how _ _ go in Saturday morning to brno* and for the part: *good day* the dialog act: *conversational_domain="frame" + speech_act="opening"* and the semantic annotation: *semantics="GREETING"* . The dialog act: *conversational_domain="task" + speech_act="request_info* and semantic annotation: *semantics="DEPARTURE(TIME, TO(STATION))"* is assigned to the rest of the sentence. The corpus contains also handcrafted word alignment (added by annotators during the corpus creation) of every Czech – SC sentence pair. For more details about the CSC corpus see [4].

3.2 Evaluation Criteria

We use the following criteria for evaluation of our experiments. The first criterion is **Sentence Error Rate** (**SER**): It is a ratio of the number of incorrect sentence translations to the number of all translated sentences. The second criterion is **Word Error Rate** (**WER**): This criterion is adopted from ASR area and is defined as the Levensthein edit distance between the produced translation and the reference translation in percentage (a ratio of the number of all deleted, substituted and inserted produced words to the total number of reference words). The third criterion is **Position-independent Word Error Rate** (**PER**): it is simply a ratio of the number of incorrect translated words to the total number

of reference words (independent on the word order). The last criterion is **BLEU** score ([5]): it counts modified n-gram precision for output translation with respect to the reference translation. A lower value of the first three criteria and a higher value of the last one indicate better i.e. more precise translation.

3.3 Decoders

We are using two different phrase-based decoders in our experiments. The first decoder is freely available state-of-the-art factored phrase-based beam-search decoder - **MOSES** ([2]). Moses can work with factored representation of words (i.e. surface form, lemma, part-of-speech, etc.) and uses a beam-search algorithm, which solves a problem of the exponential number of possible translations (due to the exponential number of possible alignments between source and target translation), for efficient decoding. The training tools for extracting of phrases from the parallel corpus are also available, i.e. the whole translation system can be constructed given a parallel corpus only. For language modeling we use the SRILM[1] toolkit.

The second decoder is our simple monotone phrase-based decoder - **SiMPaD**. The monotonicity means using the monotone reordering model, i.e. no phrase reordering is permitted. In the decoding process we choose only one alignment which is the one with the longest phrase coverage (for example if there are three phrases: p_1, p_2, p_3 coverage three words: w_1, w_2, w_3, where $p_1 = w_1 + w_2$, $p_2 = w_3$, $p_3 = w_1 + w_2 + w_3$, we choose the alignment which contains phrase p_3 only). Standard Viterbi algorithm is used for the decoding. SiMPaD uses SRILM[1] language models.

4 Experiments

4.1 Phrases Comparison

We compared the translation accuracy of handcrafted phrases with the accuracy of phrases automatically derived from the CSC corpus. The handcrafted phrases were simply obtained from the corpus. The phrase translation probability was estimated by the relative frequency ([3]):

$$\phi(\bar{s}_i|\bar{t}_i) = \frac{count(\bar{s}_i, \bar{t}_i)}{\sum_{\bar{s}_i} count(\bar{s}_i, \bar{t}_i)}$$

We used training tools of Moses decoder for acquiring the automatically derived phrases. The phrases were acquired from Giza++ word alignment of parallel corpus (word alignment established by Giza++[2] toolkit) by some heuristics (we used the default heuristic). There are many parameters which can be specified in the training and decoding process. Unless otherwise stated we used

[1] available at http://www.speech.sri.com/projects/srilm/download.html
[2] available at http://www.isi.edu/∼och/GIZA++.html

default values of parameters (for more details see Moses' documentation in [2]). The result of training is a table of phrases with five probabilities of the translation model: phrase translation probabilities $\phi(\bar{s}_i|\bar{t}_i)$ and $\phi(\bar{t}_i|\bar{s}_i)$, lexical weights $p_w(\bar{s}_i|\bar{t}_i)$ and $p_w(\bar{t}_i|\bar{s}_i)$ (for details see [3]) and phrase penalty (always equal $e^1 = 2.718$).

For comparison of results we carried out 20 experiments with various partitioning of data to the training and test set. The average results are reported in Table 1. The first column shows results of Moses decoder run on the handcrafted phrase table (phrase translation probability $\phi(\bar{s}_i|\bar{t}_i)$ only - HPH). The second column comprises results of Moses with automatically derived phrases (again phrase translation probability $\phi(\bar{s}_i|\bar{t}_i)$ only APH_PTS) and the third column contains results of Moses with automatically derived phrases and with all five translation probabilities (APH_ALL). The same language model was used in all three cases. The best results are in boldface. We used the standard sign test for the statistical significance determination. All results are given on the level of significance = 0.05.

Table 1. The translation results for comparison of handcrafted and automatically derived phrases.

	HPH	APH_PTS	APH_ALL
SER[%]	45.30 ± 2.40	$\mathbf{44.21 \pm 2.92}$	45.30 ± 2.40
BLEU	65.17 ± 1.83	67.47 ± 1.92	$\mathbf{68.77 \pm 1.72}$
WER[%]	20.74 ± 1.31	17.42 ± 1.23	$\mathbf{16.37 \pm 1.02}$
PER[%]	11.78 ± 0.77	12.01 ± 0.89	$\mathbf{11.22 \pm 0.82}$

The results show that the automatically acquired phrases have the same or better translation accuracy than the handcrafted ones. The best result we got for automatically acquired phrases with full translation model (all five translation probabilities used in decoding process APH_FULL). However, there is a difference in size of phrase tables. The table of automatically acquired phrases contains 273 226 items (phrases of maximal length 7, the whole corpus) while the table of handcrafted phrases contains 5415 items only. The size of phrase table affects a speed of translation, the smaller table the faster decoding.

4.2 "Class-based" Language Model

As well as in the area of ASR, there are problems with out-of-vocabulary words (OOV) in automatic translation area. We can translate only words which are in the translation vocabulary (we know their translation to the target language). By the analysis of the translation results we found that many OOV words are caused by missing a station or a personal name. Because the translation is limited to the domain of dialogs in train timetable information center, we decided to solve the problem of OOV words similarly as in work [6], where the class-based

language model was used for the real-time closed-captioning system of TV ice-hockey commentaries. The classes of player's names, nationalities and states were added into the standard language model in this work. Similarly, we added two classes into our language model - the class for all known station names: STATION and the class for all known personal names: PERSON. Because the semantic annotation of corpus contains station and personal names, we can simply replace these names by relevant class in training and test data and collect a vocabulary of all station names for their translation (the personal names are always spelled). Table 2 describes the results of comparison of both decoders with and without "class-based" language model. We carried out 20 experiments with a various partitioning of data to the training and test set. The standard sign test was used for statistical significance determination. The significantly better results are in boldface. In the first column there are the results of SiMPaD with a trigram language model of phrases (SiMPaD_LMP) and in the second one the results of SiMPaD with a trigram "class-based" language model of phrases (SiMPaD_CLMP). Because SiMPaD uses the table of handcrafted phrases (no more than 5.5k phrases), the used language model is based on phrases too. In the third and fourth column there are results of the Moses decoder (the phrase table of automatically acquired phrases was used) with the trigram language model (Moses_LM) and with the trigram "class-based" language model (Moses_CLM).

Table 2. The results for comparison of decoding with and without "class-based" language model.

	SiMPaD_LMP	SiMPaD_CLMP	Moses_LM	Moses_CLM
SER[%]	44.84 ± 1.96	$\mathbf{42.11 \pm 2.16}$	45.30 ± 2.40	$\mathbf{42.94 \pm 2.20}$
BLEU	67.92 ± 1.93	$\mathbf{70.68 \pm 1.73}$	68.77 ± 1.72	$\mathbf{71.17 \pm 1.69}$
WER[%]	16.02 ± 1.08	$\mathbf{14.61 \pm 0.96}$	16.37 ± 1.02	$\mathbf{15.07 \pm 1.00}$
PER[%]	13.30 ± 0.91	$\mathbf{11.97 \pm 0.80}$	11.22 ± 0.82	$\mathbf{9.94 \pm 0.77}$

The "class-based" language model is better than the standard word-based one in all cases, for both decoders and in all criteria. The perplexity of language model was reduced to about 29 % on average in the case of phrase-based models (SiMPaD) and about 28 % in the case of word-based models (Moses), from 44.45 ± 4.66 to 31.60 ± 3.38 and from 38.69 ± 3.79 to 27.99 ± 2.76, respectively. The number of OOV words was reduced to about 53 % on average for phrase-based and about 63 % for word-based models (from 1.80 % to 0.85 % and from 1.43 % to 0.54 %, respectively).

4.3 Post-processing Enhancement

We found out that for translation from the Czech to the SC we can obtain even better result when we use an additional post-processing method. Firstly, we can remove the words which are omitted in translation process (they are translated

into 'no translation' sign respectively) from the resulting translation. Anyway, to keep these words in training data gives better results (more detailed translation and language models). Secondly, we can substitute OOV words by a finger-spelling sign. Because the unknown words are finger spelled in the SC usually. The results for SiMPaD and Moses (suffix _PP for post-processing method) are in Table 3.

Table 3. The results of post-processing method in Czech $\Longrightarrow$ Signed Czech translation.

	SiMPaD_CLMP	SiMPaD_CLMP_PP	Moses_CLM	Moses_CLM_PP
SER[%]	42.11 ± 2.16	$\mathbf{40.59 \pm 2.06}$	42.94 ± 2.20	$\mathbf{41.97 \pm 2.20}$
BLEU	70.68 ± 1.73	$\mathbf{73.43 \pm 1.78}$	71.17 ± 1.69	$\mathbf{73.64 \pm 1.84}$
WER[%]	14.61 ± 0.96	$\mathbf{14.23 \pm 1.06}$	15.07 ± 1.00	$\mathbf{14.73 \pm 1.16}$
PER[%]	11.97 ± 0.80	$\mathbf{9.65 \pm 0.78}$	9.94 ± 0.77	$\mathbf{8.67 \pm 0.73}$

4.4 Czech to SC Translation

The same corpus, methods and decoders as for Czech to SC translation can be used for the inverse translation direction, i.e. from SC to Czech. The results for the SC to Czech translation are reported in Table 4. The second and the fourth columns contain results for test data where we kept also the words with 'no translation' sign and that were omitted in Czech to SC translation (suffix _WL). Finally, in the first and the third columns there are results for real test data (i.e. without the words with 'no translation' sign in the Czech to SC translation direction) (suffix _R).

Table 4. The results for Signed Czech $\Longrightarrow$ Czech translation.

	SiMPaD_CLMP_R	SiMPaD_CLMP_WL	Moses_CLM_R	Moses_CLM_WL
SER[%]	67.84 ± 1.56	57.08 ± 1.74	64.07 ± 2.71	51.74 ± 2.28
BLEU	39.55 ± 1.45	53.04 ± 1.24	50.23 ± 1.80	61.97 ± 1.80
WER[%]	36.15 ± 1.06	25.36 ± 0.78	29.65 ± 1.16	20.41 ± 1.04
PER[%]	33.04 ± 0.99	22.21 ± 0.71	26.00 ± 1.04	16.28 ± 0.85

Of course, the results for the test data containing also the words with 'no translation' sign are better, because there are more suitable words which should be in the resulting translation. Hence, for a better translation it is suitable to include the information on omitted words into the translation model. The Moses's results are better than SiMPaD's, because the word-based language model is more suitable for SC – Czech translation than phrase-based one (both trained on the corpus only).

5 Conclusion

We compared the translation accuracy of handcrafted and automatically derived phrases. The automatically derived phrases have the same or better accuracy than the handcrafted ones. However, there is a significant difference in the size of phrase tables. The table of automatically acquired phrases is more than 50 times larger than the table of handcrafted phrases. The size of phrase table affects a speed of the translation. We developed our decoder SiMPaD which uses handcrafted phrase table and some heuristics (monotone reordering and alignment with the longest phrase coverage) to speed up the translation process. We compared the SiMPaD's results with the state-of-the-art phrase-based decoder Moses. We found that the SiMPaD's results are fully comparable with the Moses's results while SiMPaD is almost 5 times faster than the Moses decoder.

We introduced "class-based" language model and post-processing method which improved the translation results from about 8.1 % (BLEU) to about 27.4 % (PER) of relative improvement in case of SiMPaD decoder and from about 7.1 % (BLEU) to about 22.7 % (PER) of relative improvement in case of Moses decoder (the relative improvement is measured between the word-based model - _LM(P) and the class-based model with post-processing - _CLM(P)_PP).

The same corpus, methods and decoders as for Czech to SC translation we used for SC to Czech translation and obtained first results for this translation direction. The experiment showed that it would be important to keep in some way the information on words that have 'no translation' sign in the Czech to SC translation direction to get better translation results.

References

1. Krňoul, Z., Železý, M., Translation and Conversion for Czech Sign Speech Synthesis, In Proceedings of 10th International Conference on TEXT, SPEECH and DIALOGUE TSD 2007. Springer-Verlag Berlin Heidelberg (2007).
2. Koehn, P. et al., Moses: Open Source Toolkit for Statistical Machine Translation. Annual Meeting of the Association for Computational Linguistics (ACL), demonstration session, Prague, Czech Republic, June 2007.
3. Koehn, P. et al., Statistical Phrase-Based Translation, HLT/NAACL, 2003.
4. Kanis, J. et al., Czech-Sign Speech Corpus for Semantic Based Machine Translation, In Lecture Notes in Artificial Intelligence, v.4188, pp.613-620, ISSN 0302-9743, 2006.
5. Papineni, K.A. et al., Bleu: a method for automatic evaluation of machine translation, Technical Report RC22176 (W0109-022), IBM Research Division, Thomas J. Watson Research Center, 2001.
6. Hoidekr, J. et al., Benefit of a class-based language model for real-time closed-captioning of TV ice-hockey commentaries, In Proceedings of LREC 2006. Paris : ELRA, 2006. s. 2064-2067. ISBN 2-9517408-2-4.