Developing a Statistical Turkish Sign Language Translation System for Primary School Students

Buse Buz
İstanbul Bilgi University
Boğaziçi University
İstanbul, Turkey
busebuzz@gmail.com

Tunga Güngör
Boğaziçi University
İstanbul, Turkey
gungort@boun.edu.tr

***Abstract*— As the access to information in the education domain increases, new technologies are developing for school children. However, deaf and dumb children still have limited access to the information, especially in their school lives. One of the most important reasons for this problem is the lack of studies in the Sign Language domain. In this paper, we propose a novel method for translation from Turkish to Turkish Sign Language for primary school students using the statistical machine translation approach. To the best of our approach, this is the first work that applies statistical translation to Turkish Sign Language. A parallel corpus is compiled from the books published by Ministry of National Education of Turkey. The results of the system were tested using different evaluation metrics. We observe that the results obtained are motivating for new studies.**

***Index Terms*— Natural Language Processing, Statistical Machine Translation, Sign Language, Turkish**

I. INTRODUCTION

Studies in recent years have showed that deaf children have encounter several problems due to their disabilities [1]. Most of them learn how to read and write in a few years while their peer groups learn it within a few months. While their peer groups can evolve their language and communication skills, deaf children can not do that because of lack of language skills and problems in their social lives. However, if they can express their thoughts and feelings with a language, they can learn a way for communication. They can also learn the written and spoken language just like their peer groups with help of the communication technique they have acquired. Sign languages are actually the communication instrument of deaf children.

The ideal aim of machine translation systems is to produce the best possible translation without human assistance. Statistical machine translation (SMT) is corpus based but slightly different in the sense that it depends on statistical modelling of the word order of the target language and of source-target word equivalences. Statistical machine translation automatically learns lexical and structural preferences from corpora. In this paper, we developed a statistical machine translation system between Turkish and Turkish Sign Language (Türk İşaret Dili - TİD). One of the biggest problems in creating such a translation system is that the number of previous studies is quite low for TİD. Furthermore, the number of competent individuals who know TİD is quite a little. Because of these reasons, it was very challenging for us to create a dataset which is one of the most important things required for a statistical translation. In this research, while building the dataset, the translation of the sign language sentences was done by us who do not know the sign language, but accompanied by supervisors who are either people who know TİD or researchers who study on TİD. Another problem is that Turkish is an agglutinative language that has a rich set of derivational suffixes and inflectional suffixes, while such attachments are not suitable for TİD. Thus, one can say that the two languages are too far away from each other in the morphological form. Therefore, for statistical translation, we also examined TİD and made some preprocessing according to the findings. Examples about the complex structure of TİD - Turkish pair can be seen in Table I. In the examples, TİD sentences are conventionally written with capital letters. Also the phrases between parentheses in TİD Sentence part do not represent TİD itself but used for representing English gloss.

TABLE I
TURKISH - TID SENTENCE PAIRS

Turkish Sentence	TID Sentence
Anne ve babası, heyecanlı olmasının doğal olduğunu söylediler. (His/her parents said it was natural to be excited.)	ANNE VE BABA (parents) SÖYLEMEK (to say) O HEYECANLI OLMAK (s/he is excited) BU NORMAL (this is normal)
Annem izin almak için okulun hangi bölümüne gitmelidir? (What part of the school should my mother go to get permission?)	BEN (I) ANNE (mother) İZİN ALMAK İÇİN (to get permission) OKUL (school) HANGİ BÖLÜM (which part) GİTMEK (to go)?
Tanımadığımız kişilerle ilişkilerimizde dikkatli olmalıyız. (We must be careful in our relations with people we don't know.)	BIZ (we) KİŞİ (someone) TANIMAKˆ DEĞİL (not to know) BIZ (we) İLİŞKİ (relation) DİKKATLİ OLMAK (to be careful) LAZIM (necessary).
Sonbaharda ağaçlar yapraklarını döker. (In the autumn the trees drop their leaves.)	SONBAHAR (autumn) AĞAÇ (tree) YAPRAK (leaf) DÖKMEK (to drop)

II. Related Works

A child's cognitive development depends on the communication and language skills. In [1], *Yorgancı et. al* already mentioned the communication problems for deaf and dumb children. To overcome this problem, the researchers created an avatar named Merry which helps deaf children to translate text using Avatar-based Interface. They set up an experiment with a test from a social studies book that was designed for primary school children. Children can read these questions by themselves, or understand the questions while watching Merry. The results show that, for deaf children, Sign Language interface has an important role. The results are shown in Table II.

TABLE II
Results with Text and Merry [1]

Accuracy	**Correct Answers**	**Wrong Answers**
Text only	45.33%	32.50%
Text and Merry	66.11%	27.08%

In [2], researchers proposed a translation system from sign language to spoken language. If we focus on the translation part, the researchers used a statistical approach instead of conventional rule-based approach. In their study, two problems have been mentioned: (1) lack of large corpora and (2) lack of a standard for notion. About the first problem, usually the corpora used for statistical translation contain about a few hundreds of thousands sentences, while there are no more than 2000 sentences in the corpus for sign languages. As for the second problem, each sign language has its own rules. Thus, every signer can show a sentence with a different way. Similar to these reasons and the number of people who know TİD is a quite limited, we also encountered the same problems while doing this research.

In the same study, to perform experiments, 1399 sentences have been used. The corpus was divided into training samples (83% of the sentences) and testing samples (17% of the sentences) [2]. Training is performed by using both IBM Model 1-4 [3] and Hidden Markov Model [4]. For evaluation metrics, mWER (word error rate) and mPER (position-independent word error rate) have been used [5]. If we consider the results in Table III, we can say that the results are promising for a statistical translation model.

TABLE III
Results for German to German Sign Language (DGS) [2]

	mWER(%)	**mPER(%)**
Single words	85.4	43.9
Alignment Templates	59.9	23.6

The most common opinion about corpus size on SMT is ”the more the better”. However, [6] shows that rule-based and statistical approaches can be compared in the sign language domain. As already mentioned, small corpus is the main problem for the statistical approach. However, in this study, corpora of different sizes were used. JRC-Acquis-L is a large corpus and JRC-Acquis-S is a small corpus drawn from the the same data. Four languages were used for translation which are from English (EN) to Romanian (RO), Romanian to English, German (GER) to Romanian and Romanian to English. The results can be seen in Table IV. If we compare BLEU [8] and TER [9] scores for different language pairs, we can see that a large data set does not make one of the scores superior to the other. With this study, we clarified that the number of sentences in a small corpus also can be sufficient to perform SMT approach.

TABLE IV
BLEU vs TER Scores

Score	**JRC-Acquis-S**	**JRC-Acquis-L**
BLEU (EN to RO)	0.4801	0.4015
TER (EN to RO)	0.5032	0.5023
BLEU (RO to EN)	0.4904	0.4255
TER (RO to EN)	0.4509	0.4457
BLEU (GER to RO)	0.2811	0.3644
TER (GER to RO)	0.6658	0.6113
BLEU (RO to GER)	0.2926	0.3726
TER (RO to GER)	0.6816	0.6112

III. Methodology

A. *Architecture*

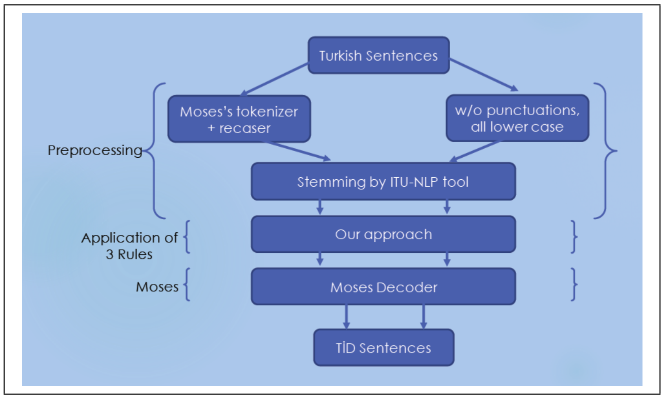

Fig. 1. System Architecture

The system consists of three steps can be seen in Fig. 1. It starts with a preprocessing part which includes tokenization, recasing and stemming. Then the proposed rules are applied. As the final step, the parallel corpora are given to the statistical machine translation system.

B. *Preprocessing*

Before training and testing our system, some processes are applied to our corpus. Tokenization indicates splitting up a sequence of strings into pieces such as words, keywords, phrases, symbols and other elements called tokens. As a first step in our study, tokenization has been applied by Moses' tokenizer [10]. After tokenization, Moses's recaser has been used. The recaser checks the first tokens of sentences to be sure whether they are starting with capital letter or not. Then the initial words in each sentence are converted to their most probable casing. In this way, data sparsity has been reduced.

After preparing the data for training the translation system, stemming is applied. Stemming is the process of reducing inflected or derived tokens to their roots. Because Turkish is an agglutinative and morphologically rich language, the aim of stemming in our study is to reduce inflectional forms of a word to a common root. Different forms of a word can be seen in Table V. Both Moses' tokenizer and ITU NLP tool [11] used for tokenization seperately. Because Turkish is not supported by Moses, ITU NLP tool was used for both tokenization and stemming to perform better results.

TABLE V
DIFFERENT FORMS OF A WORD "OKUL" (*School*)

okulun (of school)	okul (school)
okula (to school)	okul (school)
okuldan (from school)	okul (school)

TABLE VI
EXAMPLE OUTPUT FROM PREPROCESSING

FORM	LEMMA	UPOS	FEATS
Yasemin	Yasemin	Prop	-
erkenden	erken	Adv	-
kalktı	kalk	Verb	Pos$\Vert Past \Vert A3sg$
.	.	Punc	-

Yasemin erkenden kalktı.
Yasemin woke up early.

An example output after preprocessing can be seen in Table VI. All tokens have Universal Part-of-speech tags which is important for our proposed method.

C. Proposed Method

After the preprocessing is completed, some operations are applied to the sentences before feeding the parallel corpus to the Moses system. To apply such operations, rules for TİD are prepared by TİD researchers. The structures of Turkish and TİD are examined and according to the information gained, three operations are formed which are (1) adding negation, (2) adding pronoun to noun and (3) adding pronoun to verb. These operations are explained by TİD instructors. The reason for using additional operations is that there is no inflectional suffixes in TİD as mentioned before and they case a problem during statistical translation. After applying these operations, the parallel corpus is given to Moses.

- Adding Negation: In Turkish, if the verb is negative, +Neg suffix is added to the verb.

  gelmedi ⇒ gel + Verb + Neg | Past | A3sg

  (*S/he didn't come*)

  In TİD, there is no such suffix, instead DEĞİL tag is used after the verb.

  gelmedi ⇒ O GELMEK ˆ DEĞİL
- Adding pronoun to Noun: In Turkish, the possesive suffix is added to the noun.

  kalemim ⇒ kalem + Noun + P1sg

  (*my pencil*)

  In TID, again because there is no such suffix, pronoun is added to the noun.

  kalemim ⇒ BEN KALEM
- Adding Pronoun to Verb: In Turkish, personal suffixes added to the verb.

  okudum ⇒ oku + Verb + Past|A1sg

  (*I read*)

  In TID, according to the verb of the sentence, the pronoun which indicates who made the action, is added to the sentence.

  okudum ⇒ BEN OKUMAK

IV. EXPERIMENTS & RESULTS

A. Dataset

The dataset consists of Turkish-TİD sentence pairs where Turkish sentences are collected from first grade students' book of Life Science published by the Ministry of National Education of Turkey. The book consists of six units in the following order;

- Okulumuzda Hayat *(Life in our School)*
- Evimizde Hayat *(Life at Home)*
- Sağlıklı Hayat *(Healthy Life)*
- Guvenli Hayat *(Safe Life)*
- Ülkemizde Hayat *(Life in our Country)*
- Doğada hayat *(Life in Nature)*

There are total of 1950 sentences and about 13 thousand tokens. The sentences include about 1450 unigrams, 5500 2-grams, and 6650 3-grams. The sentence lengths for both corpora are shown in Fig. 2 and Fig. 3. 1500 of these sentences were used for training and 250 for development, and 200 for test.

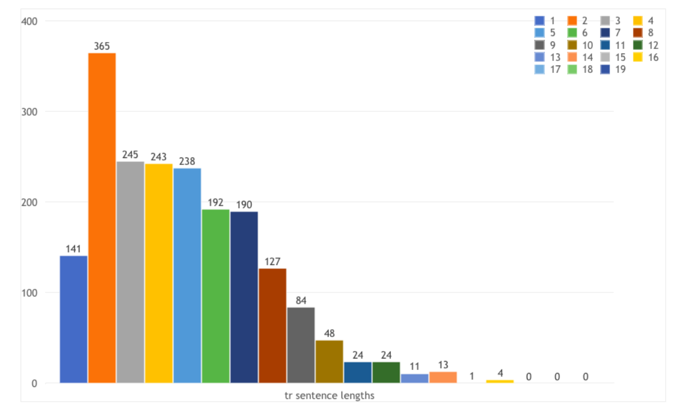

Fig. 2. Lengths of Turkish Sentences

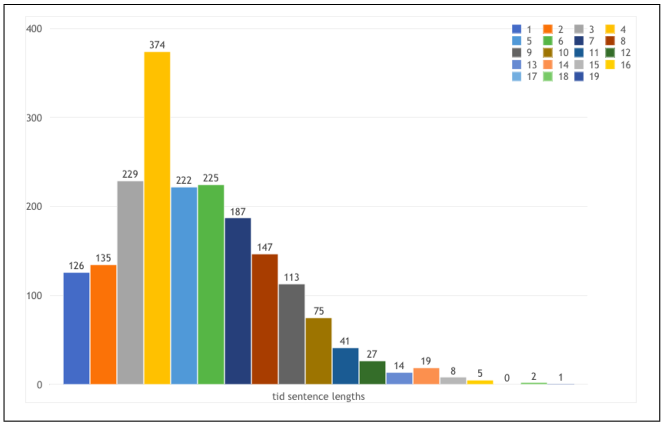

Fig. 3. Length of TID Sentences

B. Evaluation Metrics

The main evaluation metrics we used in this study are BLEU [8] and WER (word error rate) [5]. After the system outputs the translated sentences, each metric is computed with using the reference sentences. Also the metrics are used for different rates of the training and development sets.

BLEU calculates n-gram overlap between machine translation output and reference translation In Equation 1, output-length and reference length denote respectively the length of the sentences in system translation and in the reference translation. It is basically the averaged percentage of n-gram matches. For each i-gram where i = 1,2, . . . , N , it computes the percentage of the i-gram tuples in the system translation that also occurs in the reference translation (denoted as precision).

$$BLEU = min(1, \frac{output - length}{reference - length})(\prod_{i=1}^{4} precision_i)^{1/4} \tag{1}$$

WER is the minimum number of editing steps to transform output sentence to reference sentence (Equation 2). There are four possible editing steps:

- match: words match, no cost
- insertion: add word
- deletion: drop word
- substitution: replace one word with another

$$WordErrorRate = \frac{insertion + deletion + substitution}{numberofwordsinreference} \tag{2}$$

C. Results

In this study, five different approaches were examined. For each approach the data set was randomly divided into train, development and test sets. For each approach, BLEU scores can be seen in Fig. 4.

For the first approach, no operation was performed on the data. For pairs in the parallel corpus, tokenization was performed and punctuation marks were removed. After these preprocessing steps, the data were given to the Moses.

In the second approach, stemming was applied. Only the roots of tokens were considered. Post-stemming results are higher then before, because this approach eliminates data sparsity. However, the positive/negative meaning of verbs and possessive suffixes for nouns and verbs were also discarded. With this step a sharp drop can be seen in WER in Table VII

TABLE VII
SMT RESULTS

	BLEU-1	BLEU-2	BLEU-3	BLEU-4	WER
Approach 1	61.69	49.19	39.27	31.18	42%
Approach 2	77.66	64.87	54.61	46.02	30%
Approach 3	76.95	65.71	56.24	48.07	28%
Approach 4	79.63	65.86	54.22	45.23	32%
Approach 5	80.83	66.98	55.37	46.46	29%

For the first step of the proposed method, each verb were examined to find whether it has negative or positive meaning. If there is a negative tag for a verb, +Neg tag is added. By this way, negative and positive verbs do not lose their meanings after stemming. This step is called Approach 3.

In Approach 4, nouns were checked whether they had possessive suffixes or not. For these nouns, pronouns has been added. By this way, the BLEU-1 score has been increased. However, adding pronouns to any possessive suffix, caused having large number of pronouns in the sentence, and in this case BLEU-2 and BLEU-3 scores dropped.

Last step is Approach 5. In this step, verbs have been examined. If the verb has person agreement then pronoun is added to the sentence. With this method, decrease in BLEU-2 and BLEU-3 scores can be explained. [12] The pronouns from the second approach and the pronouns from the third approach led to extra tokens in the sentence. Due to the lack of fully established rules within the TİD, it is not easy to choose the pronoun for given verbs and nouns.

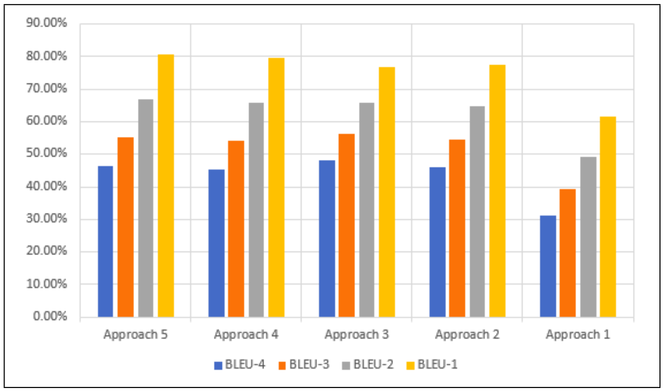

Fig. 4. BLEU scores

For this study, because there were no other study with TİD, the results were compared with the baseline model where statistical machine translation approach was not used. In the baseline results, after all the mentioned approaches applied, only word to word matches were considered. Ordering of the tokens and probabilities for n-grams were not considered.

TABLE VIII

BASELINE RESULTS

Without SMT	**BLEU-1**	**BLEU-2**	**BLEU-3**	**BLEU-4**	**WER**
Approach 1	32.73	20.79	14.40	10.26	68%
Approach 2	55.73	42.98	32.86	24.85	46%
Approach 3	54.45	43.24	34.57	27.90	48%
Approach 4	60.19	42.80	29.75	20.09	55%
Approach 5	57.65	39.79	28.82	21.31	63%

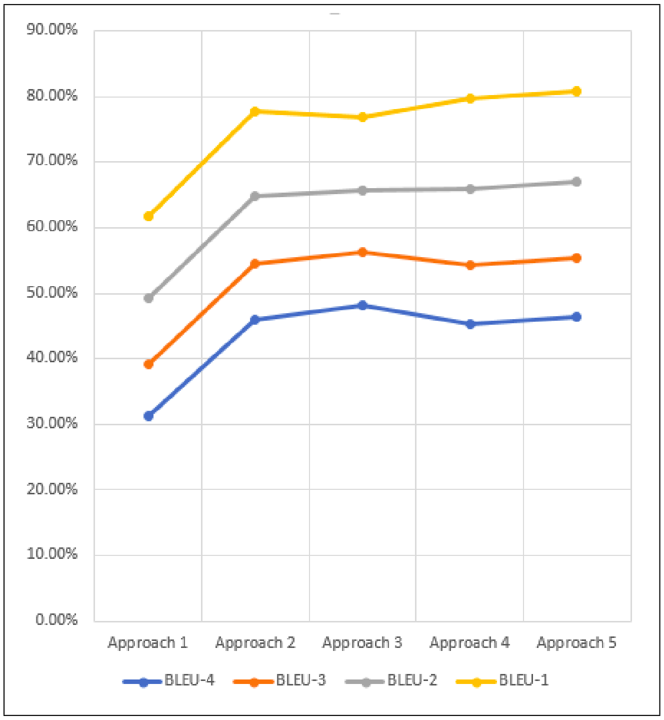

Fig. 5. Changes in BLEU scores

In Fig. 5 changes of BLEU scores can be seen. The reason why BLEU-3 and BLEU-4 scores fall after certain stage is that the system continuously adds pronouns without examining the structure and elements of the sentence because of fourth and fifth approaches. In each step, word alignments get better because every suffix can be represented in TİD. Because there are lots of pronouns coming from nouns and verbs, number of overlapping unigrams increases. However, due to the same reason, the number of overlapping 3-grams and 4-grams decreases.

This situation can be explained by an example, for each noun which has possessive suffix, pronoun ”ben” (*I*) is added to the sentence given below due to the fourth approach.

Turkish sentence: Ben, ablam, annem, babam, babaannem ve buyukbabam birlikte yasiyoruz.

(I, my sister, mother, father, grandmother and grandfather live together.)

After all processes: ben ben abla ben anne ben baba ben babaanne ve ben buyukbaba birlikte biz yasa

(I I sister I mother I father I grandmother I grandfather together we to live)

V. CONCLUSIONS

With this study, for the first time, translation from Turkish into TİD was performed by using SMT. This system also adds a new approach to the TİD studies which are quite few in the Machine Translation field. Also about 2000 new TİD translations are added to the literature. It is shown that the size of the corpus -which is thought to be the most important issue in statistical translation- is not crucial for us to apply SMT approach to a closed domain of TİD.

The system has been also tested in different situations. The approach which has relatively highest score was attempted with the 10-fold cross validation and the train/development sets of different sizes. This shows us that with more studies on TİD, different approaches can be created and the translation system can be improved.

With this study followings can be deduced,

- SMT can also be meaningful with little data.
- System performance can be improved with different approaches and data to be added in the domain.
- Such a system may be included in the translation system from Turkish written sources to the TİD visual sources.

As future work, more data and algorithms can be added to the system. Another task to do next can be adding visualization of translation for school children. Also, Neural Machine Translation can be tried for sign language translation. This way we can update our work to new era of deep learning.

ACKNOWLEDGMENT

We would like to thank to Aslı Özkul and Sumru Özsoy for their support in forming the TİD translation of Turkish sentences.

REFERENCES

[1] R. Yorganci, A. Alp Kindiroglu and H. Kose, Avatar-based Sign Language Training Interface for Primary School Education, Workshop: Graphical and Robotic Embodied Agents for Therapeutic Systems, 2016.

[2] Bungeroth, J. and H. Ney, Statistical sign language translation, Workshop on representation and processing of sign languages, LREC, Vol. 4, pp. 105-108, 2004.

[3] Brown, Peter F., et al. The mathematics of statistical machine translation: Parameter estimation. Computational linguistics 19.2 (1993): 263-311.

[4] Och, Franz Josef, and Hermann Ney. A comparison of alignment models for statistical machine translation. COLING 2000 Volume 2: The 18th International Conference on Computational Linguistics. Vol. 2. 2000.

[5] Och, F. J., Minimum error rate training in statistical machine translation, Proceedings of the 41st Annual Meeting on Association for Computational Linguistics, Vol. 1, pp. 160-167, 2003.

[6] Gavrila, M. and C. Vertan, Training Data in Statistical Machine Translation-the More, the Better?, Proceedings of the International Conference Recent Advances in Natural Language Processing 2011 , pp. 551-556, 2011.

[7] Philipp Koehn, AStatistical Machine Translation System User Manual and Code Guide, http://www.statmt.org/moses/manual/manual.pdf, 2010, (accessed at January 2019).

[8] Koehn, P., H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens et al., Moses: Open source toolkit for statistical machine translation, Proceedings of the 45th annual meeting of the ACL on interactive poster and demonstration sessions, pp. 177-180, 2007.

[9] Papineni, K., S. Roukos, T. Ward and W.-J. Zhu, BLEU: a method for automatic evaluation of machine translation, Proceedings of the 40th annual meeting on association for computational linguistics, pp. 311-318, 2002.

[10] Snover, M., B. Dorr, R. Schwartz, L. Micciulla and J. Makhoul, A study of translation edit rate with targeted human annotation, Proceedings of association for machine translation in the Americas, Vol. 200, No. 6, 2006.

[11] Eryigit, Gulsen, ITU Turkish NLP web service, Proceedings of the Demonstrations at the 14th Conference of the European Chapter of the Association for Computational Linguistics, pp. 1-4, 2014

[12] Aaron Smith, Christian Hardmeier, Jörg Tiedemann, Climbing Mount BLEU: The Strange World ofReachable High-BLEU Translations, Baltic J. Modern Computing, Vol. 4 (2016), No. 2, pp. 269–281.