A Neural Machine Translation Approach to Translate Text to Pictographs in a Medical Speech Translation System - The BabelDr Use Case

Jonathan Mutal[1] Jonathan.Mutal@unige.ch
Pierrette Bouillon[1] Pierrette.Bouillon@unige.ch
Johanna Gerlach[1] Johanna.Gerlach@unige.ch
Magali Norré[1,2] Magali.Norre@uclouvain.be
Lucía Ormaechea Grijalba[1] Lucia.OrmaecheaGrijalba@unige.ch

[1]TIM, FTI, University of Geneva, Geneva, 1205, Switzerland
[2]CENTAL, ILC, Catholic University of Louvain, Louvain-la-Neuve, 1348, Belgium

Abstract

The use of images has been shown to positively affect patient comprehension in medical settings, in particular to deliver specific medical instructions. However, tools that automatically translate sentences into pictographs are still scarce due to the lack of resources. Previous studies have focused on the translation of sentences into pictographs by using WordNet combined with rule-based approaches and deep learning methods. In this work, we showed how we leveraged the BabelDr system, a speech to speech translator for medical triage, to build a speech to pictograph translator using Unified Medical Language System (UMLS) and neural machine translation approaches. We showed that the translation from French sentences to a UMLS gloss can be viewed as a machine translation task and that a Multilingual Neural Machine Translation system achieved the best results.

1 Introduction

Patients, especially those with limited health literacy skills, often have trouble understanding health information. One of the ways in which medical communication can be facilitated is through the use of pictographs. In particular, pictographs have been used extensively to deliver specific medical instructions. The use of images has been shown to positively affect patient comprehension by improving attention, recall, satisfaction, and adherence (Houts et al., 2006; Katz et al., 2006).

Although the potential of pictographs has often been recognised, tools that automatically translate sentences into pictographs are still very scarce due to the lack of resources, which also impedes evaluation in this domain. Glyph is an automatic healthcare data processing system that automatically converts texts into sets of illustrations using natural language processing and computer graphics techniques (Bui et al., 2012). The system is based on 600 pictographs that are linked to Unified Medical Language System (UMLS, (Bodenreider, 2004)) and has been shown to have a positive impact on information recall, satisfaction, and the understandability of instructions (Hill et al., 2016). Some online medical translators also include pictographs, for example, "My Symptoms Translator" (Alvarez, 2014) and "Medipicto AP-HP", but they remain very limited in coverage and can only translate predefined sentences. There are generic MT sys-

tems that can produce pictographs (for example, Text2Picto and, more recently, PictoBERT), but they are not specialized in the medical domain. Both Text2Picto (Sevens, 2018; Vandeghinste et al., 2015) and PictoBERT (Pereira et al., 2022) are based on WordNet (Miller, 1995), which does not contain specialized medical terminology and mainly provides word-based mapping into pictographs. For example, "prise de sang" (blood draw) and "prendre le sang" (take blood), which both correspond to the same medical UMLS term (Collection of blood specimen for laboratory procedure), will each be represented by two WordNet concepts (blood + draw and take + blood) and therefore mapped to three different pictographs (Norré et al., 2022), even though the meaning is the same.

BabelDr (Bouillon et al., 2021) is a medical speech translation system specifically designed to allow French-speaking doctors to interview foreign patients in emergency settings when interpreters are not available. It can be characterised as a speech-enabled fixed-phrase medical translator and maps oral doctor interactions (questions and instructions) to a fixed set of sentences that have been pre-translated by humans, using neural machine translation methods and synthetic data. The synthetic data used to train the system are generated with a synchronous grammar that links possible source variations to the closest pre-translated sentence (called "core sentences" here). The system translates in two main phases: first, speech recognition is followed by back translation of the speech recognition result into a core sentence using neural approaches. Secondly, if this back-translated sentence is accepted by the doctor, the target sentence is produced for the patient. The system has been used since 2018 at the Geneva University Hospitals (HUG), in the context of medical dialogue with the migrant population (Janakiram et al, 2021). Translating sentences into pictographs can be another way of improving the communication between the doctor and the patient. Pictographs can also facilitate the translation process, since doctors may be able to validate a back translation into pictographs more intuitively than a core sentence that may be lexically and syntactically very different. For example, the source sentence "avez-vous envie de vomir" (do you feel like vomiting) will be back translated into "avez-vous des nausées ?" (do you have nausea) (Spechbach et al., 2017). Another advantage is the compositionality of pictographs, which enables the coverage of the system to be easily extended.

The aim of this paper is to show how we leveraged the BabelDr architecture and, in particular, the synchronous grammar to build a flexible translator from speech to pictographs for the medical domain, using synthetic data and neural MT architecture. We want to see if it is possible to build a MT system that translates doctor interactions into a semantic gloss based on UMLS concepts. This gloss defines the pictographic language, namely the concepts that are to be produced in pictographs and their syntax. Our hypotheses are that 1) the UMLS gloss is an effective way of characterizing pictographic language, 2) the mapping to UMLS gloss can be viewed as a machine translation task (see also, Mujjiga et al., 2019), and 3) a Multilingual Neural Machine Translation (Johnson et al., 2017) system that exploits both the core sentences and UMLS glosses achieves the best performance. Our contribution is twofold: on the one hand, our study allows us to compare different architectures that can translate BabelDr content into UMLS gloss and, on the other hand, it produces resources that can be shared with the community[1].

This paper is structured as follows: section 2 presents the background. This is followed by Section 3 which describes the synthetic data used to train the systems and Section 4 which outlines the translation systems. Section 5 describes the evaluation methodology, followed by results in Section 6 and conclusions in Section 7.

[1]The synthetic data used for training the systems to translate into UMLS gloss are available upon request.

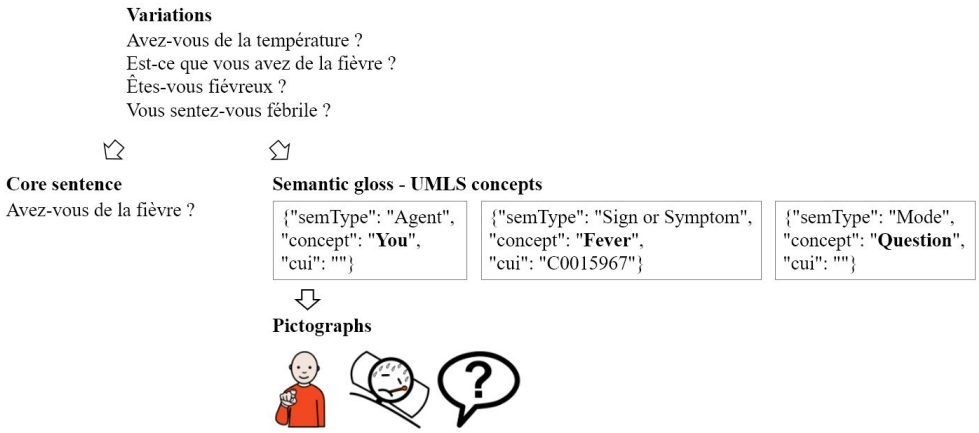

Figure 1: Overview of the steps from source variations through the semantic gloss to a sequence of Arasaac pictographs

2 Automatic translation into pictographic forms: architectures

Due to the lack of resources, translations into pictographic form are traditionally carried out using rule-based methods in three main steps. The sentence is first pre-processed and potentially simplified. Multi-word expressions are identified and lexical items are mapped to a set of disambiguated concepts. Finally, these concepts are linked to the corresponding pictographs using pictographic databases and possibly exploiting the lexical network, for example, synonyms or hyperonyms if a word has no equivalent pictograph. Existing open source databases are all based on WordNet senses (or WOLF, a WordNet equivalent for French) and link word senses to Arasaac[2] pictographs (Norré et al., 2021). In the medical Glyph system (Bui et al., 2012), the process is slightly different. Medical terminology is first identified using the UMLS ontology and images are then composed in a second step based on the semantic type and syntactic pattern.

Recently, neural methods have also been used to generate the pictographs. In particular, PictoBERT (Pereira et al., 2022) is a word-sense language representation model that predicts pictographs, also using WordNet and the Arasaac database.

In this study, we propose to translate French sentences into pictographs using NMT methods, but instead of generating the pictographs based on the WordNet word-senses, we will translate the source sentence into a semantic gloss that defines the pictographic language, namely the medical concepts to be produced in pictographs and their syntax. Like with Glyph, the gloss is based on the medical ontology and contains UMLS concepts and other linguistic elements, such as question marks and entities that are presented in a standard order based on semantic patterns (for example, <you/patient> <Sign or symptom> <time> <question>). Glossing has been used in many NLP applications, for example, in sign language MT, in which it also defines manual signs and their syntax (Ebling, 2016) and in MT of low-resourced languages (Zhou et al., 2019). For our purposes, this approach has many advantages, particularly when it comes to dealing with paraphrases of the same medical questions/instructions. For example, ”je vais prendre du sang” (I will take blood), ”je vais analyser le sang” (I will analyse your blood) and ”je vais faire une prise de sang” (I will do a blood draw) will all be mapped to the

[2]The pictographs are the property of the Aragon Government and were created by Sergio Palao for Arasaac (`https://arasaac.org`). The Aragon Government distributes them under the Creative Commons License.

```
{
   "variations": ["de la fièvre",
    "(fiévreux | fiévreuse|fébrile)",
    "$savez_vous si vous avez de la fièvre ?(en ce moment|maintenant)",
    "$êtes_vous (fiévreux | fiévreuse|fébrile) ?(en ce moment|maintenant)",
    "$sentez_vous (fiévreux | fiévreuse|fébrile) ?(en ce moment|maintenant)",
    "$avez_vous de la (fièvre | température) ?(en ce moment|maintenant)"],
   "target": {
    "frenchCoreSent": "avez-vous de la fièvre ?",
    "umlsGloss": [{"semType": "Entity","concept": "You","cui": ""},
                  {"semType": "Sign or Symptom","concept": "Fever","cui": "C0015967"},
                  {"semType": "Mode","concept": "Question","cui": "" }]
   }
}
```

Figure 2: Example of the grammar mapping source variations to core sentence and UMLS gloss

same gloss and represented by the same pictographs. This pivot representation also makes it possible to easily generate different pictographic languages, depending on the target language of the patient. The use of UMLS instead of WordNet allows us to work with medical terms instead of words. Figure 1 provides an overview of the proposed approach.

3 Grammars and data

Due to confidentiality issues, training data for spoken French medical dialogues is scarce. As previously mentioned, all BabelDr training data are generated from a manually defined Synchronous Context Free Grammar (SCFG, Aho and Ullman, 1969) that maps source variation to core sentences, using variables that are described in a formalism similar to regular expressions. This grammar was defined in close collaboration with doctors who helped collect core sentences and their possible variations. For the current project, we extended this synchronous grammar to also generate semantic glosses. Concretely, all grammar rules were manually linked to a UMLS gloss with the help of the UMLS API. Figure 2 provides an example of a rule with the different surface variations and their corresponding core sentence and UMLS gloss.

The current version of the grammar includes 2629 utterance rules that are organised by medical domain, such as abdomen, traumatology, etc., which expand into 10'991 core sentences and UMLS glosses once variables are replaced by values. These core sentences and UMLS glosses are mapped to hundreds of millions of surface variations.

4 System Settings

In this study, we experiment with two different systems trained on the synthetic data with UMLS glosses. We compare them to a baseline trained on the variations - core sentences data. In this baseline, the system translates French sentences into a core sentence and uses the grammar to generate the UMLS gloss, as in the current BabelDr architecture (see more, Mutal et al., 2019).

In this section, we explain the settings for the systems.

4.1 Data

To produce the training data, we filtered the data generated from the grammar (see Section 3) based on source language N-grams, as described in (Mutal et al., 2020). We then built two aligned corpora: one that contains source variations and the corresponding core sentences and another one with the variations and the UMLS gloss. The former is used to train a system that maps the variations to a core sentence as in Mutal et al. (2019), and the latter is used to train a system that translates into the UMLS gloss.

Variation (Source)	le mal à la tête irradie-t-il vers le haut
Core Sentence (Target)	la douleur irradie-t-elle vers le haut de la tête ?
UMLS gloss (concept names)	Pain Radiating_to Towards Upper Head Question
UMLS gloss (CUI) (Target)	C0030193 C0332301 C3875150 C1282910 C0018670 Question

Table 1: Example of data used for the training corpora. The target is created using CUI (UMLS Concept Unique Identifier) and entities (if there is no CUI).

	#Words	**#Vocabulary**
Variations	7.2M	5,105
Core Sentences	6M	2,652
UMLS Glosses	3.8M	1,667

Table 2: The number of words and vocabulary for variations, core sentences and UMLS glosses for the 746,462 sentences in the training data.

Table 1 provides an example of the data used in the training corpora. The number of words and unique words (vocabulary) of the 746,462 sentences included in the filtered set are presented in Table 2.

4.2 Systems

As we wanted to compare different systems that can translate into a UMLS gloss, we trained the models using the same settings and architecture. They were trained using the open-source implementation from OpenNMT-py (Klein et al., 2018) of Transformer architecture (Vaswani et al., 2017). Since a lot of resources are required to search the optimized hyper-parameters, we re-used the same hyper-parameters for all the models.

Baseline: Baseline system that translates variations into core sentences. This system is trained with the variations to core sentences corpus. The core sentences are then mapped to the corresponding UMLS gloss using the synchronous grammar. When the system produces output that does not match a core sentence, no UMLS gloss is produced.

UMLS NMT: System that directly translates variations into a UMLS gloss. This system is trained with the aligned corpus containing the variations and the corresponding UMLS gloss. This system produces UMLS glosses for all sentences.

Multilingual NMT: System that translates variations into both UMLS gloss and core sentence. Training a multilingual system can be beneficial to produce both UMLS and core sentences using only one model. It also helps with the training step since it shares the representation space from UMLS and core sentences (Firat et al., 2016; Kudugunta et al., 2019; Zhou et al., 2019). We trained the multilingual system using both variations to French and variations to UMLS gloss. We added a special tag at the beginning of the sentence, as shown in Table 3, to identify the target language (as suggested by, Johnson et al., 2017; Wu et al., 2021).

We extracted around 5% of the data for the development set. We verified that the variations are not in the core sentences nor variations of the training set.

5 Evaluation methodology

The aim of the system is to produce a gloss that has the same meaning as the sentence and the right syntax, so that it can produce the expected pictographic form. We carried out an automatic and human evaluation to assess the systems' performance on real medical dialogue

Source	Target
<FR>les maux migrent dans la partie basse ventre	la douleur se déplace vers la partie basse du ventre ?
<UMLS>les maux migrent dans la partie basse ventre	C0030193 C1299988 C3875150 C0230166 Question

Table 3: Training example for the MNMT system

data collected at HUG with doctors using the BabelDr system. The automatic evaluation aims at giving an overview of the systems' precision and recall at the level of concepts. The human evaluation is intended to measure whether or not the gloss expresses the same meaning as the original sentence and can therefore be used as the pivot for pictographs.

In the following sections, we present the test data, followed by the automatic and human evaluation designs.

5.1 Test Data

To estimate the quality of the systems, we used the speech data collected in real settings during a cohort study at the outpatient emergency unit of the HUG (Janakiram et al., 2020). The data were collected using the BabelDr system, which was used by doctors when interviewing real patients. The doctors were familiar with system coverage and the types of utterances to use. The data were then transcribed and manually associated with the closest core sentence. Each core sentence was then linked to its corresponding UMLS gloss using the grammar. The data consist of 883 segments, which corresponds to 5,672 words in the transcriptions (average length of 6.4 words per sentence). The following example contains an extract of a sequence of doctor utterances:

avez-vous mal à la tête maintenant ? (does your head hurt now?)

pouvez-vous me montrer avec le doigt où est la douleur ? (can you show me with your finger where the pain is located?)

depuis combien de jours avez-vous mal à la tête ? (for how many days has your head hurt?)

avez-vous déjà eu ce type de douleur ? (have you ever had this kind of pain in the past?)

avez-vous la tête qui tourne ? (is your head spinning?)

The sentences were tagged as "In Domain" if a core sentence with a similar meaning was found in the BabelDr grammars and "Out Of Domain" if not. Based on this, 92% of the sentences were "In Domain". Additionally, sentences were tagged as "In Coverage" if the variation was found in the training data (18%).

	Definition
Hypothesis	UMLS gloss generated by the system
Reference	UMLS gloss generated by the grammar for the reference core sentence
True Positives	Number of correct UMLS concepts in the hypothesis
False Positives	Number of additional UMLS concepts in the hypothesis
False Negatives	Number of missing UMLS concepts in the hypothesis

Table 4: Definition of True Positive, False Negative and False Positive.

5.2 Automatic Evaluation

To evaluate the systems, we assessed the output using Precision, Recall and F_β (Powers, 2011) on UMLS concepts, as described in Table 4. We chose $\beta = 0.5$ to give more weight to precision. As the test data is not balanced in terms of distribution of core sentences, we computed the performance for each core sentence, and then averaged it over the number of core sentences, i.e. by macro-averaging (Jurafsky and Martin, 2014). The macro-average better reflects the statistics of the less frequent core sentences and is therefore more suitable for situations in which all core sentences are equally important but are not represented equally in the test data.

5.3 Human Evaluation

We carried out a human evaluation to measure the fidelity of the UMLS glosses produced by the system. For this evaluation, we only used results from the best MT system, namely Multilingual NMT. The evaluations were carried out at the segment level by two participants.

We presented the sentences (transcriptions of doctor utterances) side by side with the system output (the UMLS gloss), in the order of dialogue. For each pair, the participants were asked to rate the UMLS gloss using one of the following categories: Same meaning, Different meaning, Related meaning or I don't know. This "Related meaning" category was to be used for cases in which the gloss only partially represented the meaning of the sentence, but could be considered to be usable in the context of the medical dialogue, for example, when one of the gloss concepts was a hyperonym or hyponym, or when the gloss contained additional information or omissions (for instance, the tense marker). We then calculated the percentage for each category. We also calculated Cohen's kappa score to measure the level of agreement between the participants.

6 Results

6.1 Automatic Evaluation

Table 5 presents the results of the automatic evaluation. The results show that the systems trained with variations to UMLS outperformed the model trained with variations to core sentences in all the metrics. For In Domain segments, Multilingual NMT outperformed all the models, but for In Coverage UMLS NMT slightly outperformed the multilingual model (0.882 vs. 0.880 on $F_{0.5}$). A closer look at the In Coverage segments revealed that the multilingual NMT sometimes added adverbs that were not present in the reference, in particular when the training includes core sentences with and without these adverbs. For example, for 'est-ce que vous toussez ?" (are you coughing?), the system UMLS NMT correctly produces "You Coughing Question", while the Multilingual NMT generates "You Coughing Very_Much Question". The reference is "You Coughing Question". In context, the adverb does not considerably affect the meaning and so both glosses may be considered to be equivalent by the doctors since they allow them to collect the same medical information, as measured through human evaluation.

	In Domain			**In Coverage**			**Out of Coverage**		
	Precision	**Recall**	**F0.5**	**Precision**	**Recall**	**F0.5**	**Precision**	**Recall**	**F0.5**
Baseline	0.788	0.80	0.78	0.844	0.857	0.844	0.77	0.785	0.77
UMLS NMT	0.814	0.828	0.814	0.883	0.886	0.882	0.793	0.811	0.796
Multilingual NMT	0.819	0.83	0.819	0.882	0.883	0.880	0.802	0.815	0.802

Table 5: Results of Automatic Evaluation.

	Same Meaning	Different Meaning	Related Meaning	I don't know	Total
All	62.47% / 65.99%	20.86% / 19.27%	14.17% / 14.74%	2.49% / 0%	882
Out Of Domain	8.45% / 14.08%	67.61% / 64.79%	21.13% / 21.13%	2.82% / 0%	71
In Domain	67.20% / 70.53%	16.77% / 15.29%	13.56% / 14.18%	2.47% / 0%	811
In Coverage	82.43% / 89.86%	6.08% / 6.08%	10.81% / 4.05%	0.68% / 0%	148
Out of Coverage	58.45% / 61.17%	23.84% / 21.93%	14.85% / 16.89%	2.86% / 0%	734

Table 6: Results of Human Evaluation for the two evaluators (eval. 1/eval. 2).

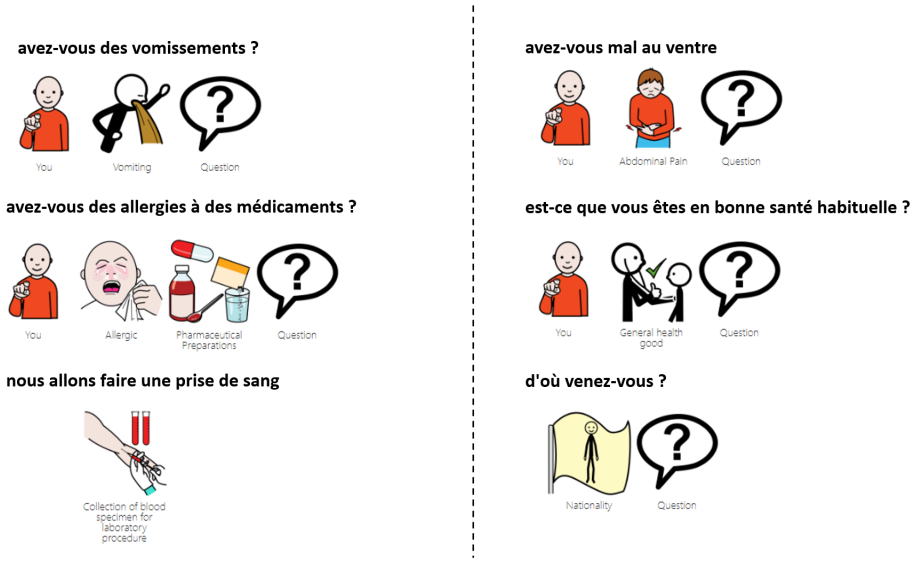

Figure 3: Examples extracted from the test data. The system translated the human transcriptions (sentences in bold) to UMLS gloss (terms below each picture). The result was then mapped to Arasaac pictographs.

6.2 Human Evaluation

Human evaluation (see Table 6) shows that the system produces an incorrect gloss for 20.1% (average for the two evaluators) of interactions, which means that four speech interactions out of five may potentially provide a correct translation into pictographs. These incorrect glosses correspond to 67.61% of Out of Domain sentences but only 16.77% of the In domain sentences. For Out of Domain sentences, there is no corresponding core sentence in the training data, but in 11.3% (Same meaning) + 21.1% of cases (Related meaning), the system was able to generalize and produce a potentially useful gloss. The agreement between the participants is 0.71 (p-value=0), which suggests that there was substantial degree of consistency in the evaluation (Landis and Koch, 1977). Some examples translated by the Multilingual system using the test data are given in Figure 3.

7 Conclusion

The aim of this paper was to propose an architecture to automatically translate doctor's interactions into pictographs. Different generic systems exist, but they are not specialised for the medical domain. The proposed method contains two steps: first, sentences are translated into a UMLS gloss based on synthetic data and secondly, concepts are mapped to pictographs. The

aim of the UMLS gloss is to characterise the pictographic language, i.e. both the medical concepts and the syntax. This paper focused on the first step.

Our contributions are twofold. On one hand, we confirmed our hypotheses and showed that mapping to the UMLS gloss can be seen as a MT task and that NMT models directly trained with UMLS glosses achieved higher F-scores. The resulting system is still limited in coverage, but results are encouraging, since 30% of Out of Domain utterances are translated into a potentially useful UMLS gloss. The human evaluation has also shown that the proposed UMLS gloss is readable by humans and can characterise pictographic language. On the other hand, this work is the first of its kind and constitutes an initial step in constituting resources (corpus and UMLS to pictographs database) and testing the impact of pictographs on medical communication. The synthetic data used for training the systems to translate into UMLS gloss are available upon request.

In the future, we plan to build upon this work in different directions. First, we plan to build a pictographic database that links UMLS concepts to pictographs, based on existing open source pictograph databases, such as Arasaac and SantéBD and other resources for the illustration of concepts (e.g. BabelNet). A preliminary study shows that only 69.7% of the BabelDr UMLS concepts can be linked to at least one Arasaac pictograph. For medical dialogues, the main problems include the representation of generic words (disease, infection, inflammation), the name of diseases (diabetes, syphilis, etc.) and temporal elements. This step will allow us to build a speech-to-pictographs baseline system for medical dialogues and will allow experiments to be carried out on the medical dialogue task itself.

More NMT architectures will be tested. Medical dialogues contain a lot of incomplete sentences (ellipsis), as explained in (Mutal et al., 2020). In the current version of BabelDr, the translation is performed in context (with the previous sentence of the dialogue) when an ellipsis is identified. Contextual NMT can also be used when translating into the UMLS gloss. UMLS semantic type and BabelDr human translations in other languages can also be added as another language in the multilingual system. The UMLS can also be used in the current architecture as an interlingua to improve translation into low-resource languages (Johnson et al., 2017).

One of our objectives is to produce more training data with more core sentences. We can easily change the synthetic data to include new core sentences. In the current version, each new grammar rule has to be translated into the 9 BabelDr target languages (Gerlach et al., 2018). The grammar therefore often groups together quasi-paraphrases, as shown in Figure 2 to reduce human translation effort (for example, "avez-vous de la fièvre" (do you have fever) and "avez-vous de la fièvre maintenant" (do you have fever now) are mapped to the same core sentence, based on the assumption that they allow doctors to collect the same anamnestic information. These rules may be split in order to only include exact paraphrases. Other resources will also be added, based on existing HUG terminological resources.

Finally, the selection of training data from the data generated by the grammar is based on N-grams in the source language (Mutal et al., 2020). We can try to select the training data based on UMLS N-grams. Our test corpus also contains a high number of In Coverage sentences, since it was collected with the BabelDr tool. We are in the process of collecting more data.

8 Acknowledgements

This work is part of the PROPICTO project, funded by the Fonds National Suisse (N°197864) and the Agence Nationale de la Recherche (ANR-20-CE93-0005).

References

Aho, A. and Ullman, J. (1969). Syntax directed translations and the pushdown assembler. *Journal of Computer and System Sciences*, 3(1):37–56.

Alvarez, J. (2014). Visual design. A step towards multicultural health care. *Arch Argent Pediatr*, 112(1):33–40.

Bodenreider, O. (2004). The unified medical language system (umls): integrating biomedical terminology. *Nucleic Acids Research*, 32(90001):267D – 270.

Bouillon, P., Gerlach, J., Mutal, J., Tsourakis, N., and Spechbach, H. (2021). A speech-enabled fixed-phrase translator for healthcare accessibility. In *Proceedings of the 1st Workshop on NLP for Positive Impact*, page 135–142, Online. Association for Computational Linguistics.

Bui, D. D. A., Nakamura, C., Bray, B. E., and Zeng-Treitler, Q. (2012). Automated illustration of patients instructions. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1158. American Medical Informatics Association.

Ebling, S. (2016). *Automatic Translation from German to Synthesized Swiss German Sign Language*. Thesis for the degree of Doctor in Philosophy, University of Zurich.

Firat, O., Cho, K., and Bengio, Y. (2016). Multi-way, multilingual neural machine translation with a shared attention mechanism. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 866–875, San Diego, California. Association for Computational Linguistics.

Gerlach, J., Spechbach, H., and Bouillon, P. (2018). Creating an online translation platform to build target language resources for a medical phraselator. In *Proceedings of the 40th edition of Translating and the Computer Conference (TC40)*, pages 60–65. AsLing, The International Association for Advancement in Language Technology.

Hill, B., Perri-Moore, S., Kuang, J., Bray, B. E., Ngo, L., Doig, A., and Zeng-Treitler, Q. (2016). Automated pictographic illustration of discharge instructions with Glyph: impact on patient recall and satisfaction. *Journal of the American Medical Informatics Association*, 23(6):1136–1142.

Houts, P. S., Doak, C. C., Doak, L. G., and Loscalzo, M. J. (2006). The role of pictures in improving health communication: A review of research on attention, comprehension, recall, and adherence. *Patient Education and Counseling*, 61(2):173–190.

Janakiram, A. A., Gerlach, J., Vuadens-Lehmann, A., Bouillon, P., and Spechbach, H. (2020). *User Satisfaction with a Speech-Enabled Translator in Emergency Settings*, pages 1421–1422. Digital Personalized Health and Medicine. IOS. ID: unige:139233.

Johnson, M., Schuster, M., Le, Q. V., Krikun, M., Wu, Y., Chen, Z., Thorat, N., Viégas, F., Wattenberg, M., Corrado, G., Hughes, M., and Dean, J. (2017). Google's multilingual neural machine translation system: Enabling zero-shot translation. *Transactions of the Association for Computational Linguistics*, 5:339–351.

Jurafsky, D. and Martin, J. H. (2014). *Speech and language processing*. Always learning. Pearson Education, 2. ed., pearson new internat. ed edition.

Katz, M. G., Kripalani, S., and Weiss, B. D. (2006). Use of pictorial aids in medication instructions: A review of the literature. *American journal of health-system pharmacy*, 63(23):2391–2397.

Klein, G., Kim, Y., Deng, Y., Nguyen, V., Senellart, J., and Rush, A. (2018). OpenNMT: Neural machine translation toolkit. In *Proceedings of the 13th Conference of the Association for Machine Translation in the Americas (Volume 1: Research Track)*, pages 177–184, Boston, MA. Association for Machine Translation in the Americas.

Kudugunta, S., Bapna, A., Caswell, I., and Firat, O. (2019). Investigating multilingual NMT representations at scale. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 1565–1575, Hong Kong, China. Association for Computational Linguistics.

Landis, J. R. and Koch, G. G. (1977). The measurement of observer agreement for categorical data. *Biometrics*, 33(1).

Miller, G. A. (1995). WordNet: A lexical database for English. *Communications of the ACM*, 38(11):39–41.

Mujjiga, S., Krishna, V., Chakravarthi, K., and J, V. (2019). Identifying semantics in clinical reports using neural machine translation. *Proceedings of the AAAI Conference on Artificial Intelligence*, 33:9552–9557.

Mutal, J., Bouillon, P., Gerlach, J., Estrella, P., and Spechbach, H. (2019). Monolingual back-translation in a medical speech translation system for diagnostic interviews - a NMT approach. In *Proceedings of Machine Translation Summit XVII: Translator, Project and User Tracks*, pages 196–203, Dublin, Ireland. European Association for Machine Translation.

Mutal, J., Gerlach, J., Bouillon, P., and Spechbach, H. (2020). Ellipsis translation for a medical speech to speech translation system. In *Proceedings of the 22nd Annual Conference of the European Association for Machine Translation*, pages 281–290, Lisboa, Portugal. European Association for Machine Translation.

Norré, M., Vandeghinste, V., Bouillon, P., and François, T. (2021). Extending a Text-to-Pictograph System to French and to Arasaac. In *Proceedings of the International Conference on Recent Advances in Natural Language Processing (RANLP 2021)*, pages 1050–1059.

Norré, M., Vandeghinste, V., Bouillon, P., and François, T. (2022). Investigating the Medical Coverage of a Translation System into Pictographs for Patients with an Intellectual Disability. In *Ninth Workshop on Speech and Language Processing for Assistive Technologies (SLPAT-2022)*, pages 44–49. Association for Computational Linguistics.

Pereira, J. A., Macêdo, D., Zanchettin, C., Oliveira, A. L. I. d., and Fidalgo, R. d. N. (2022). Pictobert: Transformers for next pictogram prediction. *Expert Systems with Applications*, 202:117231.

Powers, D. M. W. (2011). Evaluation: From precision, recall and f-measure to roc., informedness, markedness & correlation. *Journal of Machine Learning Technologies*, 2(1):37–63.

Sevens, L. (2018). *Words Divide, Pictographs Unite: Pictograph Communication Technologies for People with an Intellectual Disability*. LOT, JK Utrecht, The Netherlands.

Vandeghinste, V., Schuurman, I., Sevens, L., and Eynde, F. V. (2015). Translating text into pictographs. *Natural Language Engineering*, 23(2):217–244.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc.

Wu, L., Cheng, S., Wang, M., and Li, L. (2021). Language tags matter for zero-shot neural machine translation. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 3001–3007, Online. Association for Computational Linguistics.

Zhou, Z., Levin, L. S., Mortensen, D. R., and Waibel, A. H. (2019). Using interlinear glosses as pivot in low-resource multilingual machine translation. *arXiv: Computation and Language*.