Ellipsis Translation for a Medical Speech to Speech Translation System

Jonathan Mutal[1], **Johanna Gerlach**[1], **Pierrette Bouillon**[1], and **Hervé Spechbach**[2]

[1]FTI/TIM, University of Geneva, Switzerland
[2]Hôpitaux Universitaires de Genève (HUG), Switzerland
{Jonathan.Mutal, Johanna.Gerlach, Pierrette.Bouillon}@unige.ch
herve.spechbach@hcuge.ch

Abstract

In diagnostic interviews, elliptical utterances allow doctors to question patients in a more efficient and economical way. However, literal translation of such incomplete utterances is rarely possible without affecting communication. Previous studies have focused on automatic ellipsis detection and resolution, but only few specifically address the problem of automatic translation of ellipsis. In this work, we evaluate four different approaches to translate ellipsis in medical dialogues in the context of the speech to speech translation system BabelDr. We also investigate the impact of training data, using an undersampling method and data with elliptical utterances in context. Results show that the best model is able to translate 88% of elliptical utterances correctly.

1 Introduction

Ellipsis is one of the least studied discursive phenomena in automatic translation. Like anaphora, ellipsis require context to be understood, but contrary to anaphora, there is no indicator that there is a missing part in the sentence[1]. The characterising feature of ellipsis is that "elements of semantic content are obtained in the absence of any corresponding form. The syntax thus appears to be incomplete. More specifically, the implicit semantic context is recovered from elements of linguistic and extralinguistic context" (Ginzburg and Miller, 2018).

In NLP, different studies have focused on automatic ellipsis detection and resolution either with rules (patterns or grammars) (for example, the pioneer work from Hardt, 1992) or classification techniques (for example, Hardt and Rambow, 2001; Bos and Spenader, 2011; Liu et al., 2016; Kenyon-Dean et al., 2016; McShane and Babkin, 2016; Rønning et al., 2018). However, only few studies specifically address this problem in machine translation (MT), despite the recent interest for context modelling in neural machine translation (see for example, Bawden et al., 2018). Very recently, some qualitative studies showed the negative impact of ellipsis on generalist neural systems (DeepL, Google Translate, etc.) from a translation point of view in the English-French pair (for example, Hamza, 2019).

In this paper, we focus on automatic translation of ellipsis in medical dialogues, in the particular context of BabelDr, a speech to speech translation system for the medical domain (Spechbach et al., 2019)[2]. Elliptical utterances are very common in dialogues, since they ensure the principle of economy and provide a way to avoid duplication (Hamza et al., 2019). In the medical dialogues we are interested in, ellipsis allows doctors to question patients in a more efficient way (Where is your pain? In the back? Is the pain severe? Moderate?) (Tanguy et al., 2011). Literal translation of these elliptical utterances is rarely possible without affecting communication, in particular with structurally different languages which do not share the same type of ellipsis. For example in Japanese, adjectival ellipsis are very informal and should be

[1] Ellipsis is "a case of anaphora, where the anaphor is a null proform (zero-anaphora)" (Ginzburg and Miller, 2018)

[2] https://babeldr.unige.ch/

translated by complete sentences (Bouillon et al., 2007). The following examples illustrate elliptical utterances where literal translation is problematic, as it produces agreement errors, wrong prepositions or other syntactical or grammatical issues that can make the elliptical utterance difficult to understand.

```
Source: is the pain intense?
->MT: la douleur est-elle intense
Source: sudden?
-> MT: ⋆soudain

Source: do you have pain in
your stomach?
-> MT: le duele el estómago?
Source: in your head?
-> MT: ⋆en la cabeza?

Source: is the pain severe
-> MT: hageshii itami desu ka?
Source: moderate?
-> MT: ⋆chuuteido?
```

The aim of this paper is to compare different approaches to translate ellipsis in the context of BabelDr. Section 2 describes the BabelDr system. Section 3 outlines the methodology, including the objective and research questions, the test data and the evaluation metrics. Section 4 presents the approaches and models, followed by Section 5 which describes the different sets of training data. Section 6 presents the results and Section 7 concludes.

2 The context: BabelDr

2.1 The BabelDr system

BabelDr is a speech-enabled fixed-phrase translator designed to allow French speaking doctors to carry out diagnostic interviews with patients with whom they don't have any common language in emergency settings where no interpreters are available. It combines speech recognition with manually pre-translated sentences, grouped by diagnostic domains. Doctors can freely speak their questions, the system maps the recognised utterance (hereafter: *variation*) to the closest pre-translated sentence (hereafter: *core sentence*), and, after approval by the doctor, the core sentence is translated for the patient. This ensures the reliability of speech recognition and of translation, essential for safe use in the medical domain.

The scarcity of training data available for this domain, a consequence of data confidentiality issues and of the minority languages involved (e.g., Tigrinya, Farsi, Albanian), has at first led to the development of a grammar-based approach. A Synchronous Context Free Grammar (SCFG, Aho and Ullman, 1969) which describes source language variation patterns and their mapping to core sentences is used to compile a language model used by Nuance for speech recognition. This grammar based speech recognition produces high quality results for in coverage items. To handle sentences that are out of grammar coverage, BabelDr also includes a large vocabulary recogniser. Results from this recogniser must then be mapped to the closest core sentences, a task to which several approaches have been applied, including tf-idf indexing and dynamic programming (DP, Rayner et al., 2017) and, more recently, a NMT approach (Mutal et al., 2019). The latter is one of the approaches evaluated in the present study, where it has been extended to handle elliptical utterances.

2.2 Ellipsis in BabelDr

In the BabelDr context, instead of producing a literal translation of the ellipsis, we aim at mapping elliptical utterances to the closest non-elliptical core sentence, for which translations are available in the system. This presents the advantage of removing all ambiguity related to ellipsis and their translation. To resolve the ellipsis, we use context information, which in a diagnostic interview is the previous translated utterance. The proposed ellipsis processing workflow is illustrated in Figure 1 and will be discussed in further detail in Section 4.

3 Methodology

3.1 Objective and research question

The aim of this study is to evaluate the performance of four different approaches for the ellipsis translation task: indexing, classification, neural machine translation and hybrid.

The research questions guiding our experiments are listed as follows 1) What is the best approach to handle ellipsis in this context? 2) How does the distribution of class instances affect the performance of the proposed models? 3) Does inclusion of ellipsis-specific training data improve performance?

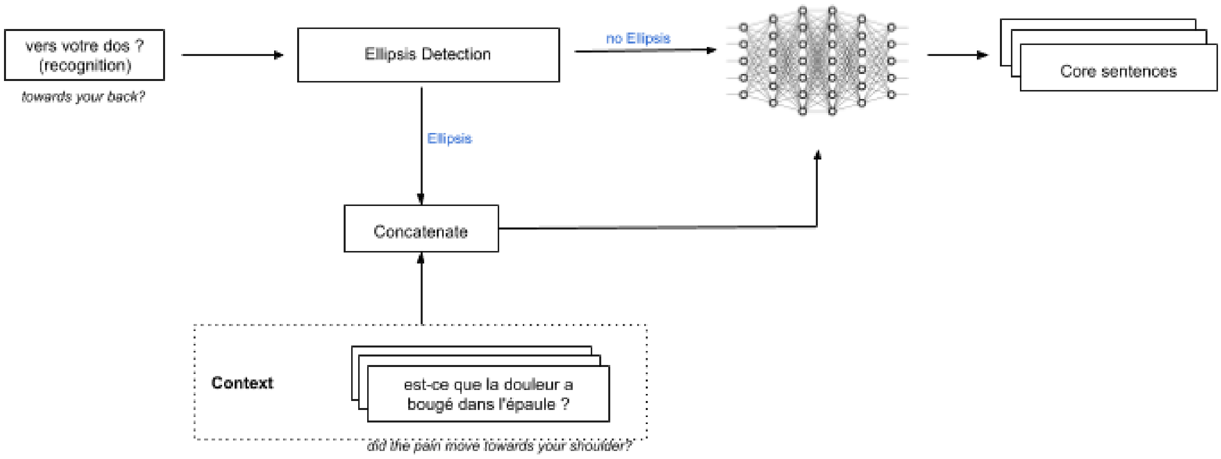

Figure 1: Ellipsis translation task in BabelDr: overview.

3.2 Test data

Since the currently deployed version of babelDr only handles ellipsis in a limited manner, doctors were instructed to use only complete sentences. Consequently, real usage data contains very few elliptical utterances. For this study, we have therefore used a test suite based on the BabelDr coverage and described in (Rayner et al., 2018). This was created by extracting the list of available core sentences for the abdominal domain and transforming complete sentences into elliptical sentences where possible, for example:

```
avez-vous mal au ventre
avez-vous mal dans le bas-ventre
    --> dans le bas-ventre
avez-vous mal dans le haut du
ventre
    --> le haut du ventre
```

Each elliptical utterance was associated with a corresponding complete utterance to serve as context. Five native francophone subjects were then asked to speak the pairs (context and elliptical utterance) in a natural way, freely varying the wording, but with the instruction to respect the distinction between elliptical and plain utterances. Data were collected using a web tool which prompted the subjects and recorded their responses. This produced a total of 1'676 recorded utterances. Each utterance was then transcribed and matched to the most plausible core sentence by two judges and when necessary disagreement between judges resolved. If the second sentence of the pair was not elliptical because subjects did not follow instructions, they were removed from the test suite. This process finally produced 838 recorded pairs, with the corresponding core sentences. The average utterance length was 8.96 words for the plain utterances and 3.14 words for the elliptical utterances (Rayner et al., 2018).

Since the focus of this study is not on speech recognition performance, but on the subsequent processing, we performed our experiments with the transcriptions as input rather than the speech recogniser output, thereby assuming recognition is perfect.

3.3 Evaluation

We want to compare the different approaches at the task level, namely how many elliptical utterances will result in a correct translation for the patient. Since the system relies on human pre-translation (cf. Section 2), a correct core sentence is equivalent to a correct translation. We therefore measured the sentence error rate (SER), defined as the percentage of utterances for which the resulting core sentence is not identical to the annotated correct core sentence.

Since the target is a finite set of sentences, we also measured system performance on the test data using three standard metrics for classification: recall, precision and F1. As the test data is not perfectly balanced, we computed the performance for each class, and then averaged over the number of classes, i.e. by macro-averaging. The macro-average better reflects the statistics of the smaller classes and therefore is more appropriate when all classes are equally important (Jurafsky and Martin, 2014). We could have applied the standard BLEU score for the evaluation of the MT approaches, but since it is not applicable to the other approaches, it is not appropriate for our comparison.

The metrics were calculated using a module in Sklearn[3].

4 Approaches

As mentioned earlier, our objective is to use the context (previous utterance) to map elliptical utterances to the closest core sentence. Figure 1 provides an overview of the ellipsis translation task as it would be performed in BabelDr. Starting with a source sentence, we perform ellipsis detection using a binary classifier (support-vector machine) trained on handcrafted features. In this context, elliptical sentences can easily be detected by sentence length and syntactic structure. Therefore, the sentence length, the first word of the sentence and its part-of-speech are used as features to train the classifier. This method achieves 98% of accuracy on ellipsis detection. If the utterance is identified as an ellipsis, it is concatenated with the previous utterance from the dialog (Tiedemann and Scherrer, 2017). This concatenated sentence is then processed like other utterances.

In the following sections, we describe the four approaches applied after concatenation. The same training data (described in Section 5) was used for all approaches. The source sentences were preprocessed using the same method for all the models: they have been lower cased and tokenized. Each approach has its own built-in tokenization method to reach optimal results, except for machine translation where we applied BPE.

4.1 Indexing

In this approach, the task is to find the source variations that are the closest matches for a new utterance. To do so, each sentence was represented by a vector and a similarity metric was used to compare them. We employed two approaches to embed each sentence:

tf-idf The first approach uses a customised tf-idf (Salton and Buckley, 1988), where tf-idf was applied to subword occurrences (two to four characters) in variations for a given core sentence. Common pre-processing methods for tf-idf are lemmatizing and removing stop words; however, since accurate preservation of meaning is imperative in a medical dialog context, e.g. in terms of verb tenses, in our experiments words were left as word forms.

Universal Sentence Encoder The second approach uses the current state-of-the-art for multilingual encoding (Chidambaram et al., 2019). To encode each source sentence, we used an already trained Universal Sentence Encoder[4] (hereafter *uencoder*).

We then used the approximate nearest neighbor search (Andoni and Indyk, 2006) to extract the closest variation sentence with cosine similarity, and return the corresponding core sentence.

4.2 Sequence Classification

In this approach, the task is to classify each variation into a core sentence using a distance based classification method (Xing et al., 2010). We trained two different neural classifiers:

CamemBERT This classifier uses the current state-of-the-art for French based on RoBERTa (Liu et al., 2019), which is used for many NLP tasks. We used the CamemBERT pre-trained model (Martin et al., 2019) and added a classification layer on top of the model to fine-tune it with our data (Sun et al., 2019). To do so, we set-up 10 epochs using the Transformer framework for python (Wolf et al., 2020).

fastText The second approach uses a sequence classification baseline based on bag of tricks (Joulin et al., 2017). We used fastText on bigrams with 100 epochs and a learning rate of 0,2. The other hyper parameters were set by default[5].

4.3 Machine Translation

With these approaches, the task is to translate the source utterance into a core sentence. We have trained two different NMT models:

LSTM We trained a neural machine translation model with an embedding size of 512 in the encoder and decoder. Encoder and decoder were each composed of two LSTM (Hochreiter and Schmidhuber, 1997) with an attention mechanism on the decoder side (Bahdanau et al., 2014; Luong et al., 2015). The model was trained with a dropout rate of 0.3 and a batch size of 64 examples. This system is described in detail in (Mutal et al., 2019).

[3] https://scikit-learn.org/stable/modules/generated/sklearn.metrics.classification_report.html

[4] https://tfhub.dev/google/universal-sentence-encoder-multilingual/3

[5] https://fasttext.cc/docs/en/supervised-tutorial.html

Corpus	Subset	#sentences	#words	#vocabulary
All data	Train	21M	322M	3121
	Dev	2M	35M	2923
Sampled	Train	143’011	1.5M	3’095
	Dev	15’891	176’816	2’413
Ellipsis Corpus	Train	394’767	4.7M	3’218
	Dev	43’863	528’175	2’593

Table 1: Number of sentences, words and vocabulary on source variations for each training data.

Ellipsis	Core sentence
aux épaules (shoulders)	la douleur se déplace vers les épaules ? (the pain moves towards the shoulders?) la douleur au ventre irradie-t-elle vers les épaules ? (does the belly pain radiate to the shoulders?) avez-vous aussi mal aux épaules ? (do your shoulders hurt as well?)
du foie (liver)	avez-vous eu un examen du foie ? (have you had a liver exam?) avez-vous eu un contrôle médical du foie ? (have you had a liver checkup?) avez-vous un cancer du foie ? (do you have liver cancer?)

Table 2: Two examples of ellipsis with corresponding possible core sentence

Transformer The second model relies on a transformer based architecture for machine translation (Vaswani et al., 2017) with default parameters and size[6].

For both architectures, early stopping was used to reduce the number of training steps by monitoring the performance on the development set. We used OpenNMT (Klein et al., 2017) to train the models.

4.4 Hybrid

The hybrid approach combines the best neural machine translation model with the best classification model to build an N-best list of sentences, in this experiment a 2-best list which includes the core sentence generated by machine translation and one sentence from the classification results. To select the best result in this list, we used the log probability of the generated core sentence from the neural machine translation: if it was below a threshold (< -0.25), we kept the core sentence generated by the classifier, else we kept the NMT result. The threshold was set based on the observation that 93% of the sentences above that threshold were mistranslated.

5 Training Data

In this section, we describe the training data sets used for this study. All data were generated from a recent version of the BabelDr SCFG for the abdominal diagnostic domain and consist of variation-core pairs. Table 1 summarises the number of sentences, words and vocabulary for each set.

5.1 All Data

The main data set includes 23M variations, of which 321’698 are ambiguous (i.e. sentences that can be mapped to more than one core sentence). Most of these ambiguous sentences are elliptical. Table 2 shows two examples of such sentences. The variations are mapped to the 4’132 different core sentences available for the abdominal domain. These core sentences are not represented equally in the corpus: 50% of the 4’132 core sentences occur less than 52 times in the data. For example, the core sentence “avez-vous pris des médicaments contre la douleur ?” (have you taken any painkillers?) is mapped to 3’496’503 source variations (14% of the entire dataset) whereas “avez-vous de l’oxygène à la maison ?” (do you have oxygen at home?) is only mapped to 5 source variations.

Since we are interested in evaluating the complete set of core sentences, we have maintained the same distribution when splitting the data into development and training.

[6] https://opennmt.net/OpenNMT-tf/model.html

	Source variation	Core sentence
Generated from grammar		
Context	la douleur a bougé dans l'épaule ? (did the pain move to your shoulder?)	la douleur se déplace vers les épaules ? (does the pain move to your shoulders?)
Ellipsis	vers votre dos ? (towards your back?)	la douleur se déplace vers le dos ? (the pain moves towards your back?)
Concatenated for training		
Ellipsis	la douleur a bougé dans l'épaule vers votre dos ? (did the pain move to your shoulders towards your back?)	la douleur se déplace vers le dos ? (the pain moves towards your back?)

Table 3: Example of generated ellipsis training data, composed of variation-core pairs : one complete (context), followed by a corresponding elliptical utterance. For training, elliptical variations are concatenated with the preceding variation (context)

5.2 Sampled Data

As mentioned in the previous section, our main corpus is highly imbalanced. In this context, where all core sentences are relevant for the task, the exclusion or misclassification/translation of minority categories (in our case, core sentences) on the dataset could lead to a heavy cost (Haixiang et al., 2017). Therefore, we used resampling techniques to rebalance the sample space in order to alleviate the effect of the skewed class distribution on the learning process. We applied under-sampling, which is suggested as the best alternative when the training sample size is too large (Mazurowski et al., 2008; Haixiang et al., 2017).

To reduce the number of variations by core sentence while keeping data as representative as possible, we propose a new algorithm for under-sampling based on bigrams consisting in the following steps:

1. For each core sentence, extract all bigrams present in the associated variations.
2. Build a new list of variations by iteratively extracting variations from a list in randomised order until all bigrams are covered.

After under-sampling, the resulting corpus contained 159'902 variations and 87 ambiguous samples. Furthermore, 75% of the core sentences were mapped to less than 32 variations. Even though we managed to reduce most of the categories, minority classes were still under-represented compared to the majority classes. For example, "avez-vous mal au ventre en position de chien de fusil ?" (do you have abdominal pain in a fetal position?) still had 731 variations whereas "combien de kilos avez-vous pris ?" (how much weight did you gain?) had only 1.

5.3 Ellipsis Corpus

To generate training data for ellipsis in context, we exploit grammar rules that contain variables. These variables are placeholders that are replaced by different values at system-compile time, e.g. "avez-vous pris [des anti-douleurs|des medicaments contre l'acidité|...] récemment ?" ("Did you take [painkillers|antacids|...] recently?"). To produce elliptical utterances, we have kept only the value of the variable as source variation, associated with a corresponding complete core sentence. Each of these elliptical variation-core pairs follows a matching complete variation-core pair which serves as context, as shown in Table 3.

To train the models, we transformed the elliptical source variations by concatenating them with the context source variation. The same concatenation was performed on the test data.

6 Results

In this section we first describe the evaluation of the under-sampling method (subsection 6.1). We then give results for different models trained with under-sampled data (subsection 6.2). Finally, including only the best model for each approach in terms of F1, we evaluate the impact of training on Ellipsis data (subsection 6.3).

6.1 Under-sampling

To evaluate the under-sampling method, we ran the experiment with two approaches, machine translation (LSTM, Transformer) and classification (fastText), trained with two different data sets: under-sampled data (hereafter *sampled*) and all data. We then compared performance by calculating SER, precision, recall and F1. Table 4 shows the results on test data.

Model	Data	SER	Precision	Recall	F1
LSTM	all data	0.29	0.56	0.54	0.56
	sampled	0.28	0.60	0.63	0.59
Transformer	all data	0.32	0.55	0.61	0.56
	sampled	0.30	0.58	0.62	0.57
fastText	all data	0.32	0.54	0.55	0.52
	sampled	0.29	0.56	0.57	0.55

Table 4: Models trained with under-sampled (sampled) and all training data (all data).

Approach	Model	SER	Precision	Recall	F1
Indexing	tf-idf	0.53/0.39	0.34/0.51	0.32/0.47	0.32/0.47
	uencoder	0.62/0.49	0.27/0.39	0.23/0.39	0.23/0.37
Classification	fastText	0.52/0.29	0.32/0.56	0.28/0.57	0.28/0.55
	CamemBERT	0.44/0.23	0.41/0.66	0.39/0.71	0.39/0.66
Machine Translation	LSTM	0.53/0.28	0.34/0.60	0.30/0.63	0.30/0.60
Hybrid	LSTM + CamemBERT	0.23/0.17	0.54/0.75	0.50/0.77	0.50/0.74

Table 5: Results on elliptical utterances/all on under-sampled training data for different models on indexing, classification, machine translation and hybrid.

We observe that the proposed under-sampling method (fastText-sampled, LSTM-sampled and Transformer-sampled) produces better results in this particular context indicating that a more balanced data set improves performance in terms of SER, precision, recall and F1.

Regarding the machine translation approaches, while results suggest that both architectures are suitable for the task, we observe that LSTM-sampled and LSTM slightly outperform Transformer and Transformer-sampled on SER, precision, recall and F1. Because of training data size and number of parameters, training time was considerably lower for the LSTM architecture with sampled data. Accordingly, we carried out the subsequent experiments using the LSTM model for the machine translation approach.

6.2 Approaches

In order to select the best approach and model to handle ellipsis in this context, we measured the performance of two different models for each approach (cf. section 4), except for machine translation where we already chose LSTM (cf. subsection 6.1).

Table 5 presents the SER, precision, recall and F1 for elliptical and all sentences.

Classification, with CamemBERT, achieves the best scores across all approaches for both elliptical and all sentences. For elliptical sentences only, tf-idf is the second best approach with 0.53, 0.34, 0.32, 0.32 for SER, precision, recall and F1. However, LSTM outperforms tf-idf for all sentences, showing that LSTM is better suited for non-elliptical sentences.

Based on the observation that sentences that were not well classified by CamemBERT were classified correctly by LSTM, we decided to combine LSTM and camemBERT to build a hybrid system. This hybrid achieved 0.23 and 0.50 on elliptical sentences for SER and F1, outperforming the best model by 0.21 and 0.11 for those metrics respectively. For those sentences that the hybrid classifies/translates adequately, 52% are well translated/classified by both models, 20% by LSTM only and the rest by CamemBERT only.

6.3 Ellipsis Training Data

To determine if the inclusion of ellipsis data in the training data affects performance, we selected the three best models based on the results described in the previous section and trained them with the ellipsis corpus described in section 5.3 in addition to the sampled training data. Table 6 shows final results for each model.

Results show that training models with elliptical sentences improves performance in terms of SER, precision, recall and F1. CamemBERT trained

Approach	Model	SER	Precision	Recall	F1
Classification	CamemBERT	0.15/0.08	0.75 /0.84	0.73/0.85	0.73/0.84
Machine Translation	LSTM	0.23/0.15	0.60/0.71	0.57/0.71	0.57/0.70
Hybrid	CamemBert + LSTM	**0.12**/0.06	0.78/0.86	0.77/0.87	0.76/0.86

Table 6: Results on elliptical utterances/all with ellipsis corpus added to training data.

with the additional ellipsis corpus outperforms the one trained with only the sampled data by 0.29, 0.34, 0.34 and 0.34 for each metric respectively.

With the additional ellipsis training data, Hybrid also outperforms the other approaches (88% of elliptical utterances are translated correctly), yet the difference is not as large as with plain training data only (cf. Table 5). We observed that 85% of the elliptical sentences were well classified by both models. 11% of the sentences were classified correctly by CamemBERT and badly by LSTM, and 4% the other way around.

Closer investigation of the 15% of elliptical sentences which were badly classified revealed several cases. Some of the classification errors were due to ambiguous cases where more than one core sentence would be appropriate for a given elliptical utterance. We also observe many cases where the core sentence was very close to the correct one, but more or less generic.

With these results, we confirmed that in this context, training models with ellipsis improves performance in terms of SER, precision, recall and F1.

7 Conclusion

In this study we have applied different approaches to an ellipsis translation task, in the context of a medical speech translator. We have also experimented with different forms of training data generated from the BabelDr SCFG. Results show that under-sampling the training data improves results for all tested approaches. Of all the tested systems, the hybrid approach, combining neural machine translation and classification models is the most successful both in terms of our task specific metric (SER) and in terms of precision/recall/F1. We also observe that the inclusion of ellipsis training data further improves results.

One limitation of this study is the annotation of the test data. Each source variation has been annotated with a single correct core sentence, but this does not reflect the real use case: the purpose of BabelDr is to allow doctors to collect information from the patient, not to translate their exact utterance. Often, even if the core sentence is not an exact match (e.g. “”in the lower part” vs “in the lower part of the abdomen”), in context it still allows the doctor to obtain the required information. In future work, a more task-oriented annotation approach would be interesting.

A further aspect worth investigating is exploring novel architectures to add the context in different ways: train a context aware decoder to correct translations (Voita et al., 2019, for neural machine translation,) or train a dual-source BERT (Correia and Martins, 2019) adding context on the tuning step for sequence classification.

Finally, future work will also include the replication of these experiments with data from real diagnostic interviews and with data from other diagnostic domains.

Acknowledgements

This project is funded by the ”Fondation Privée des Hôpitaux Universitaires de Genève”.

References

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

Andoni, A. and Indyk, P. (2006). Near-optimal hashing algorithms for approximate nearest neighbor in high dimensions. In *2006 47th Annual IEEE Symposium on Foundations of Computer Science (FOCS’06)*, page 459–468.

Bahdanau, D., Cho, K., and Bengio, Y. (2014). Neural machine translation by jointly learning to align and translate. *3rd International Conference on Learning Representations, ICLR 2015 - San Diego, United States.* arXiv: 1409.0473.

Bawden, R., Sennrich, R., Birch, A., and Haddow, B. (2018). Evaluating discourse phenomena in neural machine translation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, page 1304–1313. Association for Computational Linguistics.

Bos, J. and Spenader, J. (2011). An annotated corpus for the analysis of vp ellipsis. *Language Resources and Evaluation*, 45(4):463–494.

Bouillon, P., Rayner, E., Starlander, M., and Santaholma, M. E. (2007). Les ellipses dans un système de traduction automatique de la parole. In *Actes de TALN/RECITAL*, pages 53–62. ID: unige:3452.

Chidambaram, M., Yang, Y., Cer, D., Yuan, S., Sung, Y., Strope, B., and Kurzweil, R. (2019). Learning cross-lingual sentence representations via a multi-task dual-encoder model. In *Proceedings of the 4th Workshop on Representation Learning for NLP (RepL4NLP-2019)*, pages 250–259, Florence, Italy. Association for Computational Linguistics.

Correia, G. M. and Martins, A. F. T. (2019). A simple and effective approach to automatic post-editing with transfer learning. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3050–3056, Florence, Italy. Association for Computational Linguistics.

Ginzburg, J. and Miller, P. (2018). Ellipsis in head-driven phrase structure grammar. In van Craenenbroeck, J. and Temmerman, T., editors, *The Oxford Handbook of Ellipsis*, page 74–121. The Oxford Handbook of Ellipsis. Oxford University Press.

Haixiang, G., Yijing, L., Shang, J., Mingyun, G., Yuanyue, H., and Bing, G. (2017). Learning from class-imbalanced data: Review of methods and applications. *Expert Systems with Applications*, 73:220–239.

Hamza, A. (2019). *La détection et la traduction automatiques de l'ellipse : enjeux théoriques et pratiques*. PhD thesis, Université de Strasbourg STRASBOURG.

Hardt, D. (1992). An algorithm for VP ellipsis. In *30th Annual Meeting of the Association for Computational Linguistics*, pages 9–14, Newark, Delaware, USA. Association for Computational Linguistics.

Hardt, D. and Rambow, O. (2001). Generation of vp ellipsis: a corpus-based approach. In *Proceedings of the 39th Annual Meeting on Association for Computational Linguistics - ACL '01*, page 290–297. Association for Computational Linguistics.

Hochreiter, S. and Schmidhuber, J. (1997). Long short-term memory. *Neural Comput.*, 9(8):1735–1780.

Joulin, A., Grave, E., Bojanowski, P., and Mikolov, T. (2017). Bag of tricks for efficient text classification. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, page 427–431. Association for Computational Linguistics.

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

Kenyon-Dean, K., Cheung, J. C. K., and Precup, D. (2016). Verb phrase ellipsis resolution using discriminative and margin-infused algorithms. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, page 1734–1743. Association for Computational Linguistics.

Klein, G., Kim, Y., Deng, Y., Senellart, J., and Rush, A. (2017). Opennmt: Open-source toolkit for neural machine translation. In *Proceedings of ACL 2017, System Demonstrations*, page 67–72. Association for Computational Linguistics.

Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., and Stoyanov, V. (2019). Roberta: A robustly optimized bert pretraining approach. *arXiv:1907.11692 [cs]*. arXiv: 1907.11692.

Liu, Z., Gonzàlez Pellicer, E., and Gillick, D. (2016). Exploring the steps of verb phrase ellipsis. In *Proceedings of the Workshop on Coreference Resolution Beyond OntoNotes (CORBON 2016)*, page 32–40. Association for Computational Linguistics.

Luong, T., Pham, H., and Manning, C. D. (2015). Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421, Lisbon, Portugal. Association for Computational Linguistics.

Martin, L., Muller, B., Suárez, P. J. O., Dupont, Y., Romary, L., de la Clergerie, V., Seddah, D., and Sagot, B. (2019). Camembert: a tasty french language model. *arXiv:1911.03894 [cs]*. arXiv: 1911.03894.

Mazurowski, M. A., Habas, P. A., Zurada, J. M., Lo, J. Y., Baker, J. A., and Tourassi, G. D. (2008). Training neural network classifiers for medical decision making: The effects of imbalanced datasets on classification performance. *Neural Networks*, 21(2–3):427–436.

McShane, M. and Babkin, P. (2016). Detection and resolution of verb phrase ellipsis. *Linguistic Issues in Language Technology – LiLT, Volume 13, Issue 1*, page 36.

Mutal, J. D., 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 for Machine Translation, E. A., editor, *Proceedings of Machine Translation Summit XVII Volume 2: Translator, Project and User Tracks*, pages 169–203. ID: unige:123138.

Rayner, E., Gerlach, J., Bouillon, P., Tsourakis, N., and Spechbach, H. (2018). Handling ellipsis in a spoken medical phraselator. In Dutoit T., Martín-Vide C., P. G., editor, *Statistical Language and Speech Processing. SLSP 2018*, pages 140–152. Springer. ID: unige:110589.

Rayner, E., Tsourakis, N., and Gerlach, J. (2017). Lightweight spoken utterance classification with cfg, tf-idf and dynamic programming. In Camelin, N., Estève, Y., and Martín-Vide, C., editors, *Statistical Language and Speech Processing*, page 143–154. Springer International Publishing.

Rønning, O., Hardt, D., and Søgaard, A. (2018). Sluice resolution without hand-crafted features over brittle syntax trees. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, page 236–241. Association for Computational Linguistics.

Salton, G. and Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. *Inf. Process. Manage.*, 24(5):513–523.

Spechbach, H., Gerlach, J., Mazouri Karker, S., Tsourakis, N., Combescure, C., and Bouillon, P. (2019). A speech-enabled fixed-phrase translator for emergency settings: Crossover study. *JMIR Medical Informatics*, 7(2). ID: unige:117081.

Sun, C., Qiu, X., Xu, Y., and Huang, X. (2019). How to fine-tune bert for text classification? In Sun, M., Huang, X., Ji, H., Liu, Z., and Liu, Y., editors, *Chinese Computational Linguistics*, page 194–206. Springer International Publishing.

Tanguy, L., Fabre, C., Ho-Dac, L.-M., and Rebeyrolle, J. (2011). Caractérisation des échanges entre patients et médecins : approche outillée d'un corpus de consultations médicales. *Corpus, 10 |2011*, pages 137–154.

Tiedemann, J. and Scherrer, Y. (2017). Neural machine translation with extended context. In *Proceedings of the Third Workshop on Discourse in Machine Translation*, pages 82–92, Copenhagen, Denmark. Association for Computational Linguistics.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, , 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, *31st Conference on Neural Information Processing Systems (NIPS 2017), Long Beach, CA, USA*, page 5998–6008. Curran Associates, Inc.

Voita, E., Sennrich, R., and Titov, I. (2019). When a good translation is wrong in context: Context-aware machine translation improves on deixis, ellipsis, and lexical cohesion. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1198–1212, Florence, Italy. Association for Computational Linguistics.

Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., and et al. (2020). Huggingface's transformers: State-of-the-art natural language processing. *arXiv:1910.03771 [cs]*. arXiv: 1910.03771.

Xing, Z., Pei, J., and Keogh, E. (2010). A brief survey on sequence classification. *SIGKDD Explor. Newsl.*, 12(1):40–48.