Monolingual backtranslation in a medical speech translation system for diagnostic interviews - a NMT approach

Jonathan Mutal[1], **Pierrette Bouillon**[1], **Johanna Gerlach**[1], **Paula Estrella**[2], and **Hervé Spechbach**[3]

[1]FTI/TIM, University of Geneva, Switzerland
[2]FaMaF y FL, University of Córdoba, Argentina
[3]Hôpitaux Universitaires de Genève (HUG), Switzerland
{Jonathan.Mutal, Pierrette.Bouillon, Johanna.Gerlach}@unige.ch
paula.estrella@unc.edu.ar
herve.spechbach@hcuge.ch

Abstract

BabelDr is a medical speech to speech translator, where the doctor has to approve the sentence that will be translated for the patient before translation; this step is done using monolingual backtranslation, which converts the speech recognition result into a core sentence. In this work, we model this step as a simplification task and propose to use neural networks to perform the backtranslation by generating and choosing the best core sentence. Results of a task-based evaluation show that neural networks outperform previous versions of the system.

1 Introduction

BabelDr[1] is a joint project between the Faculty of Translation and Interpreting of the University of Geneva and Geneva University Hospitals (HUG) (Bouillon et al., 2017; Boujon et al., 2017).

The aim of the project is to build a speech to speech translation system for emergency settings which meets three criteria: reliability, data security and portability to low-resourced target languages relevant for the HUG. To ensure reliability, the system is based on a set of manually pre-translated sentences (around 30'000 *core sentences*) defined with the help of doctors and classified by anatomic domains (e.g. head, chest, abdomen, etc.). The basic idea is that the doctor can speak freely and the system will map the recognised utterance to the closest core sentence.

The translation from source recognition result to target language is done in two steps: 1) mapping of the source recognition result to a core sentence (*backtranslation*, Gao et al., 2006; Seligman and Dillinger, 2013) and 2) look-up of the (human) translation of the core sentence for the relevant target language.

Backtranslation is therefore an essential step in this type of architecture (see also Ehsani et al., 2008; Seligman and Dillinger, 2013). The doctor has to approve the backtranslation of his utterance, ensuring awareness of the exact meaning of the translation produced for the patient. Backtranslation can also be considered as a type of simplification task (Cardon, 2018). It translates the doctor's questions for the layman, reducing the vocabulary by 40%, removing medical jargon and making the meaning explicit both for the human translator and the patient. The following are examples of such lexical, syntactic and semantic simplification processes:

- Recognition result: *c'est chaud* (it is warm) → Backtranslation: *la peau est-elle chaude ?* (is the skin warm?)
- Recognition result: *où est-ce que se trouve la douleur* (where is the pain) → Backtranslation: *pouvez-vous me montrer avec le doigt où est la douleur ?* (can you show with your finger where the pain is?)
- Recognition result: *avez-vous un hématome* (do you have a hematoma) → Backtranslation: *avez-vous un bleu ?* (do you have a

© 2019 The authors. This article is licensed under a Creative Commons 4.0 licence, no derivative works, attribution, CC-BY-ND.

[1]More information available at https://babeldr.unige.ch/

bruise?)

In the current version of the system, backtranslation is performed by rule-based methods and methods borrowed from information retrieval. In this paper, we investigate a backtranslation approach using neural machine translation (NMT) trained on the data generated from the existing grammar. Our aim is to see whether it is possible to bootstrap the NMT from the rule-based system and how it will perform in comparison with the existing strategies used in BabelDr.

Section 2 describes BabelDr and the different strategies used for backtranslation in the current system. We then explain how NMT was derived from the grammar to create different neural network versions (Section 3). Section 4 describes the task-based evaluation and Section 5 presents the results.

2 BabelDr versions

The current BabelDr application used at the HUG translates from French to Arabic, Albanian, Farsi, Spanish, Tigrinya and French Swiss Sign Language. It is a hybrid system which combines rule-based and tf-idf methods for backtranslation. In this section we describe these different methods and the system versions used in our study.

2.1 Version 1 - rule-based version

The rule-based version of the system relies on a manually written grammar, using a formalism based on Synchronous CFG (SCFG, Aho and Ullman, 1969). The grammar consists of a set of rules defining source language variation patterns which are mapped to core sentences (Gerlach et al., 2018). This grammar is compiled into a language model which can be used by Nuance[2] for speech recognition and parsing to core sentences. While this rule based approach works well for in coverage (IC) spoken utterances, i.e. utterances that are among the variations described in the grammar, it often fails for out-of-coverage (OOC) ones. For the abdominal domain (one out of 13 diagnostic domains), the grammar currently contains 1'797 rules which map 4'082 core sentences to 488 million variations.

2.2 Version 2 - tf-idf/DP version

The second version of the system uses a large vocabulary speech recogniser (Nuance Transcription Engine) customised with data derived from the grammar. It then applies an approach based on tf-idf indexing and dynamic programming (DP) to match the recognition result to a core sentence (Rayner et al., 2017). This version is better suited for processing of OOC utterances, but remains imperfect, in particular because it relies on a bag of words approach.

2.3 Version 3 - hybrid version

The third version of the system, which is the currently deployed version, combines the rule-based method (Version 1) with the tf-idf/DP approach (Version 2) in order to benefit from the precision of the rules on IC sentences while ensuring robustness on OOC data. The results from the two methods are combined as follows: when the rule based recogniser confidence score is over a given threshold, Version 1 is used; when it is below the threshold, suggesting poor recognition, the tf-idf/DP result is used instead.

In the next sections we describe how we used NMT for backtranslation and present the experiments carried out to compare the different approaches.

3 NMT for backtranslation

As mentioned, backtranslation is seen here as a translation to a simplified language, where many variations of the same source sentence are translated into a predefined easy-to-understand core sentence. Even if simplification is a well studied process, only few studies apply machine translation and NMT (Wang et al., 2016). The main reason is the lack of aligned corpora as mentioned in (Suter et al., 2016), in particular in the medical domain and for French (Cardon, 2018). In this study, we propose to use data generated from the grammar to construct an aligned corpus and train a NMT system. The backtranslation is performed by NMT and the final result is chosen among the N-Best translations according to a heuristic (Section 3.3). In the next sections, we describe the generated corpus, explain how we trained the NMT system and introduce two BabelDr versions based on NMT.

3.1 Data set

For this experiment, we used the data generated from an early version of the SCFG, described in (Rayner et al., 2017). It consists of 221'819

[2]https://www.nuance.com

Source variation	Backtranslation
votre ventre fait mal ?	avez-vous mal au ventre ? (do you have stomach pain?)
ça vous soulage de rester couché	la douleur au ventre diminue-t-elle quand vous restez couché ? (does the stomach pain decrease when you lie down?)
avez-vous des antécédents chirurgicaux au niveau de l'abdomen ?	avez-vous eu une opération du ventre ? (have you had abdominal surgery?)
est ce que vous pourriez me montrer votre carte d'assuré ?	pouvez-vous me montrer la carte d'assurance ? (could you show me your insurance card?

Table 1: Examples of aligned sentences derived from rules (source variations-backtranslation).

sentences from the abdominal diagnostic domain mapped to 2'517 different core sentences. Table 1 illustrates examples of the data.

Since we are interested in evaluating the complete set of core sentences, development and test data follow the same distribution as the training data, i.e. each subset contains an equal proportion of core sentences. Tables 2 and 3 summarise the number of sentences, tokens and vocabulary for each subset, for source variations and core sentences (target) respectively.

Subset	#sentences	#tokens	#vocabulary
Train	199k	2M	2132
Dev	12k	124k	1581
Test	10k	103k	1478

Table 2: Number of sentences, tokens and vocabulary for source variations.

Subset	#sentences	#tokens	#vocabulary
Train	199k	1.5M	880
Dev	12k	99k	838
Test	10k	82k	829

Table 3: Number of sentences, tokens and vocabulary for core sentences (target).

The source sentences have been lower cased and tokenized; then, Byte-pair encoding (Sennrich, 2016) was trained on the training data set and applied to training, development and test data.

3.2 NMT configuration

We used OpenNMT-tf (Klein et al., 2017, OpenNMT,) for training and decoding. OpenNMT is a framework mainly focused at developing encoder-decoder architectures.

As we can consider our task a low resource NMT (2M tokens in training data, Zoph et al., 2016), we had two alternatives to tackle this task: 1) follow (Zoph et al., 2016) and apply transfer learning or 2) choose an appropriate neural architecture in terms of size. We find 2) a better alternative because of the lack of medical corpora suitable for this application.

Transformer (Vaswani et al., 2017) is the state-of-art in most NMT tasks, but it is better suited to learn in high-resource conditions (Tran et al., 2018). Therefore, we decided to compare Transformer performance with an encoder-decoder architecture based on recurrent neural networks (RNN) (Kalchbrenner and Blunsom, 2013; Bahdanau et al., 2014; Loung et al., 2015).

Transformer: The model is composed of a 512 embedding size in the encoder and decoder. The architecture is described in (Vaswani et al., 2017). The parameters used were the default for this model[3].

RNN: The model is composed of 512 embedding size in the encoder and decoder. Encoder and decoder are each composed of two LSTM (Hochreiter et al., 2006) with an attention mechanism on the decoder side (Bahdanau et al., 2014; Loung et al., 2015). The model was trained with a dropout rate of 0.3 and a batch size of 64 examples.

Both models use early stopping in order to reduce the number of training steps by monitoring the performance on the development set. All the models are trained using ADAM optimiser (Kingma and Ba, 2014). The parameters were averaged from the last 10 checkpoints for each model.

[3] http://opennmt.net/OpenNMT-tf/model.html#catalog

Speech rec. result	Avez-vous des animaux
1-best NMT	travaillez-vous avec des animaux ? (is core sentence = true)
2-best NMT	avez-vous des animaux ? (is core sentence = false)
Result	travaillez-vous avec des animaux ?

Figure 1: Example of utterance where the 1-best NMT result is a core sentence and is therefore chosen as final result

Speech rec. result	Avez-vous des nausées les vomissements
1-best NMT	vomissez-vous des boissons alcoolisées ? (is core sentence = false)
2-best NMT	vomissez-vous des nausées ? (is core sentence = false)
Closest core to 1-best	buvez-vous des boissons alcoolisées tous les jours ? (0.43)
Closest core to 2-best	avez-vous des nausées ? (0.84)
Result	avez-vous des nausées ?

Figure 2: Example of utterance where neither of the NMT results is a core sentence and final result is selected based on cosine similarity.

3.3 N-Best sentence

The model was configured to generate *n* candidates ($n = 1, 2, 3$ for this experiment); the best candidate is selected by keeping the first one which matches a core sentence. This case is illustrated in Figure 1. If none of the candidates are core sentences, the *word embedding similarity* selection heuristic from STS 2016 (see Agirre et al., 2016) is used to find the closest core sentence. In order to find the closest sentence, sentence embeddings (Arora et al., 2016) are computed using word embeddings learned by the decoder. Afterwards, the candidates (i.e. the n results generated by NMT) are embedded to the same continuous space and cosine similarity is calculated to choose the closest core sentence. Figure 2 illustrates this case.

3.4 NMT Evaluation

We carried out an automatic evaluation to choose between the two neural MT architectures, adding N-Best sentence generation to each model. We measured system performance on the test data using two standard metrics: BLEU (Papineni et al., 2002) and TER (Snover et. al, 2006), as shown in Table 4.

Model	N-Best	TER	BLEU
RNN	1-best	0.8	97.84
	2-best	0.7	99.7
	3-best	0.7	99.7
Transformer	1-best	0.9	97.65
	2-best	0.8	99.45
	3-best	0.8	99.45

Table 4: Comparison between models with N-Best (N=1,2,3) sentences.

Table 4 shows that there was no significant difference between the results obtained with Transformer and with RNN. An intuitive explanation for this is that the sentences in our data set are rather short, with a mean sentence length of 10.37 words, and thus present no difficulties for the RNN approach. Furthermore, the amount of training data might not be suitable for a transformer architecture (Tran et al., 2018). We also observe that adding the 2nd best sentence improves the performance of the model while adding a 3rd best does not bring an improvement.

To carry out the next experiments, we chose RNN with 2-best sentences.

3.5 BabelDr NMT versions (Version 4 and 5)

Two new versions of BabelDr were built based on the neural architecture described in previous Sections.

Version 4: uses the same large vocabulary speech recogniser as Version 2, but instead of an approach based on tf-idf and dynamic programming (DP), it is based on a neural approach.

Version 5: is hybrid, following the same principle as Version 3 but using NMT instead of tdf-idf to generate the core sentences when the rule-based recogniser confidence score is below the threshold.

4 Task-based evaluation

4.1 Motivation

Our main research question is to see if it is possible to bootstrap a NMT system from the data generated with the rule-based system. To answer this, we will focus on the following sub-questions: 1)

	Speech			Text		
Version	**IC**	**OOC**	**ALL**	**IC**	**OOC**	**ALL**
Version 1	13.9	72.0	31.2	0	100	29.8
Version 2	8.5	48.1	20.4	1.2	43.5	13.8
Version 3	6.4	48.1	18.8	–	–	–
Version 4	9.3	32.7	16.3	0.8	21.0	6.8
Version 5	6.2	32.2	13.9	–	–	–

Table 5: SER for IC, OOC and ALL for in domain speech recognition results (Speech) and transcriptions (Text). No text results are provided for the hybrid versions (3 and 5), since transcriptions are independent from the speech recogniser confidence score threshold.

will the system be able to generate core sentences, 2) does a non core sentence indicate an out-of-domain (OOD) utterance, i.e. one that could not be associated with any of the core sentences, and 3) how will the system perform in comparison with the currently used approaches. In order to answer these questions, we used the different versions of the system (described in Sections 2 and 3.5) to process utterances collected during diagnostic interviews. These test data are the same as used in Rayner et al. (2017). Results for system Versions 1-3 are therefore taken from this publication.

4.2 Test Data

The test data are French utterances collected in an experiment where doctors and medical students used the system to diagnose two standardised patients (Bouillon et al., 2017). It includes 10 complete diagnostic interviews by 10 different speakers, for a total of 827 utterances. Each utterance was transcribed and annotated, where possible, with a corresponding core sentence. We excluded out-of-domain (OOD) utterances, which represent 110 items (14%). The remaining data can be split into IC (503 items), where transcriptions are among the variations described in the SCFG, and OOC (214 items), where the transcriptions are not among these variations, but match a core sentence closely enough to be considered synonymous.

4.3 Evaluation criteria

We want to compare the different versions at the task level, namely how many spoken utterances will result in a correct translation for the patient. Since the system relies on human pre-translation (Section 1), 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. As input utterances we used the speech recognition results from the large vocabulary recogniser (speech) and the transcriptions (text, which simulates the case where recognition is perfect). This metric and approach allows us to compare our results with those reported for system Versions 1-3 in Rayner et al. (2017).

5 Results

In order to answer our first research question, we calculated the proportion of non core sentences among the sentences generated by the NMT system. Considering all data (IC, OOC and OOD), these only amount to 2% on 2-Best and 5% on 1-Best. Nearly 50% of these non core sentences are translations of out of domain utterances. These results suggest that non core sentence backtranslations could serve as indicator for out of domain utterances, a fact that could be exploited in the BabelDr application to identify concepts not covered by the system.

Table 5 presents SER results on test data both on speech recognition results and on transcriptions. For spoken data, the NMT model (Version 4) outperforms all the previous versions on ALL data for the task, reducing the SER by 4 points in comparison with the best of the previous versions. A closer comparison of the two non-hybrid versions shows that Version 4 has a slightly higher error rate than Version 2 on IC utterances (9.3 vs 8.5), while it has a much lower error rate on OOC utterances (32.7 vs 48.1). These results could be explained by the different approaches: since tf-idf matches words and computes its scores based on grammar content, it has more chances of finding correct results for IC utterances than NMT, which generates a new sentence based on a semantic representation. On the other hand, NMT is better suited to handle OOC, since this semantic representation allows it

to generalise.

As expected, the hybrid NMT version (Version 5) obtains similar performance to Version 4 on OOC and improves scores on IC data (6.2 vs 9.3), since as with the previous hybrid system (Version 3) the generally reliable high-confidence rule-based results replace potentially incorrect NMT results.

When using transcriptions as input, the proportion of errors for NMT is reduced by 9.5 SER points (16.3 to 6.8 on ALL data for Version 4), showing the negative impact of speech recognition errors on the result. A closer look at the data shows that most errors occur when the speech recognition result contains 1) words that are not in the training data, which often happens when words are recognised incorrectly by the large vocabulary recogniser, resulting in OOD items, or 2) words that appear in the grammar but are rare in the training data.

6 Conclusion

The results of this study show that for this backtranslation task, NMT outperforms previous versions of the system. It also shows the potential of NMT and hybrid architectures for simplification tasks.

For BabelDr, the neural network approach reduces the error by 4 SER points on spoken utterances and by 9.5 points on transcriptions, which simulate perfect speech recognition. Results also show that this approach has generated core sentences in all but 2% of cases (2-Best), suggesting that it can learn the simplified language. Non core sentences mostly indicate OOD utterances.

This study has several limitations. It uses only a subset of the sentences generated by the SCFG for training, thus allowing for words present in the rules, but missing from the training data; this is subject to further improvements by enlarging the training corpus.

Another limitation is that for this study we used an older version of the grammar. The latest version of the grammar not only includes more words (nearly 5000 for abdominal domain), core sentences and variations but also contains ambiguous rules. These rules allow multiple backtranslations for ambiguous utterances, for example *est-elle forte* (is it severe?) could translate to *la fièvre est-elle forte* (is the fever high?) or *la douleur au ventre est-elle forte ?* (is the abdominal pain severe) depending on the context, where context can be defined as the utterances before, e.g. *avez-vous de la fièvre* (do you have a fever?) for the example above. Integrating context dependent processing is thus another area for improvement of the backtranslation process. One possibility for this could be to use document-level machine translation (Lesly et al., 2018) or add the context when translating (Agrawal et al., 2018).

A further aspect worth investigating is the size of the grammar: the current grammar extensively describes variations, necessary for grammar-based speech recognition, yet it is unclear whether such an extensive grammar is necessary for the generation of training data for the NMT approach, or whether a more compact grammar, combined with the NMT approach in a hybrid system, could achieve similar performance.

Finally, future work will also include a comparison of the NMT approach with state-of-the-art approaches for semantic text similarity (STS) tasks (Zhao and Vogel, 2002; Cer et al., 2017; Rychalska et al., 2016).

Despite these limitations, to the best of our knowledge, it is the first experiment to use NMT for backtranslation in fixed phrase translators and to test it on data from real diagnostic interviews.

Acknowledgements

This project is financed by the "Fondation Privée des Hôpitaux Universitaires de Genève". We would also like to thank Nuance Inc for generously making their software available to us for research purposes.

References

Agirre Eneko, Banea Carmen, Cer Daniel, Diab Mona, Gonzalez-Agirre Aitor, Mihalcea Rada, Rigau German and Wiebe Janyce. 2016. SemEval-2016 Task 1: Semantic Textual Similarity, Monolingual and Cross-Lingual Evaluation. *Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016)*. San Diego, California. 497–511.

Agrawal, Ruchit AND Turchi, Marco AND Negri, Matteo. 2018. *Proceedings of the 21st Annual Conference of the European Association for Machine Translation*: 28-30 May 2018, Universitat d'Alacant, Alacant, Spain, pp. 11-20.

Aho, Alfred and Ullman, Jeffrey. 1969. Properties of syntax directed translations. *Journal*

of Computer and System Sciences. 3. 319–334. 10.1016/S0022-0000(69)80018-8.

Bahdanau, Dzmitry, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural Machine Translation by Jointly Learning to Align and Translate. In Proceedings of the International Conference on Learning Representations (arXiv:1409.0473).

Bouillon P, Gerlach J, Spechbach H, Tsourakis N, Halimi. 2017. BabelDr vs Google Translate: a user study at Geneva University Hospitals (HUG). Proceedings of the 20th Annual Conference of the European Association for Machine Translation. Prague, Czech Republic. p 747-52.

Boujon V, Bouillon P, Spechbach H, Gerlach J, Strasly I. 2017 September 15-16. Can speech-enabled phraselators improve healthcare accessibility? A case study comparing BabelDr with MediBabble for anamnesis in emergency settings. Proceedings of the 1st Swiss Conference on Barrier-free Communication. Winterthur, Switzerland. p. 32-38. 2017. 2018 DOI 10.21256/zhaw-3000.

Cardon, Rémi. 2018. Approche lexicale de la simplification automatique de textes médicaux. *In: RJC 2018, 14-18 May 2018.* Rennes, France.

Daniel M. Cer, Mona T. Diab, Eneko Agirre, Iñigo Lopez-Gazpio and Lucia Specia. 2017. SemEval-2017 Task 1: Semantic Textual Similarity - Multilingual and Cross-lingual Focused Evaluation. *dblp computer science bibliography, https://dblp.org. CoRR.* arXiv. 1708.00055. http://arxiv.org/abs/1708.00055. Mon, 13 Aug 2018 16:45:59 +0200.

Ehsani, Farzad, Jim Kimzey, Elaine Zuber, Demitrios Master, and Karen Sudre. 2008. Speech to Speech Translation for Nurse Patient Interaction. *In: Coling 2008: Proceedings of the Workshop on Speech Processing for Safety Critical Translation and Pervasive Applications.* Manchester, England, August, 2008, pages 54-59.

Felix Hill, Kyunghyun Cho and Anna Korhonen 2016 Learning Distributed Representations of Sentences from Unlabelled Data. *CoRR*.abs/1602.03483 .http://arxiv.org/abs/1602.03483. arXiv. Mon, 13 Aug 2018. dblp computer science bibliography, https://dblp.org.

Gao Y, Gu L , Zhou B, Sarikaya R, Afify M, Kuo K, Zhu W, Deng Y, Prosser C, Zhang W, Besacier L. IBM MASTOR SYSTEM: Multilingual automatic speech-to-speech translator. Proceedings of the First International Workshop on Medical Speech Translation, in conjunction with NAACL/HLT. 2006 June 4-9; New York, NY, USA. Madison, WI; Omnipress Inc; 2006.

Gerlach J, Spechbach H, Bouillon P. 2018 Creating an online translation platform to build target language resources for a medical phraselator. *Proceedings of the 40th Edition of the Translating and the Computer Conference (TC40).* 2018 15-16 November; London, UK. 2018.

Hochreiter S and Schmidhuber J. 2006. Long Short-Term Memory *Neural Computation, 9(8):1735–1780, 1997.*

Kingma, D. P. and Ba, J. 2014. Adam: A method for stochastic optimization. *arXiv preprintarXiv:1412.6980.*

Ke M. Tran, Arianna Bisazza and Christof Monz. 2017. The Importance of Being Recurrent for Modeling Hierarchical Structure. *dblp computer science bibliography, https://dblp.org*. Mon, 13 Aug 2018 16:46:56 +0200.

Klein, Guillaume, Yoon Kim, Yuntian Deng, Jean Senellart, and Alexander M. Rush. 2017. Opennmt: Open-source toolkit for neural machine translation. *arXiv preprintarXiv:1701.02810.*

Lesly Miculicich Werlen and Dhananjay Ram and Nikolaos Pappas and James Henderson. 2018. Document-Level Neural Machine Translation with Hierarchical Attention Networks. *CoRR*.abs/1809.01576. dblp computer science bibliography, https://dblp.org.

Luong, Thang, Hieu Pham, and Christopher D. Manning. 2015. Effective Approaches to Attention-based Neural Machine Translation. *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing.* 1412–1421.

Nal Kalchbrenner and Phil Blunsom. 2013. Recurrent Continuous Translation Models. *In Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing,* pages 1700–1709, Seattle,Washington, USA. Association for Computational Linguistics.

Papineni, Kishore, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic eval-uation of machine translation. *In*

Proceedings of the 40th annual meeting on association for computational linguistics, pages 311–318. Association for Computational Linguistics.

Rayner M, Tsourakis N, Gerlach J. 2017. Lightweight Spoken Utterance Classification with CFG, tf-idf and Dynamic Programming. *In: Camelin N., Estève Y., Martín-Vide C. (eds) Statistical Language and Speech Processing. SLSP 2017.* Lecture Notes in Computer Science, vol 10583. Springer, Cham

Rychalska, B., Pakulska, K., Chodorowska, K., WojciechWalczak and Andruszkiewicz, P. 2016. Samsung Poland NLP team at SemEval-2016 task 1 : Necessity for diversity ; combining recursive autoencoders, wordnet and ensemble methods to measure semantic similarity. *In SemEval-2016, pp. 614–620* 497–511.

Sanjeev Arora, Yingyu Liang, and Tengyu. Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings. *In ICLR 2017.*

Seligman M, Dillinger M. 2013. Automatic speech translation for healthcare: some internet and interface aspects. *Proceedings of the 10th International Conference on Terminology and Artificial Intelligence (TIA-13).* 2013 October 28-30; Paris, France. 2013.

Sennrich, Rico, Barry Haddow, and Alexandra Birch. 2016. Neural machine translation of rare words with subword units. *In Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics, ACL 2016,* August 7-12, 2016, Berlin, Ger-many, Volume 1: Long Papers.

Snover, M., B. Dorr, R. Schwartz, L. Micciulla and J.Makhoul. 2006. A study of translation edit rate withtargeted human annotation. *Proceedings of the Association for Machine Translation in the Americas,* Vol. 200, No. 6.

Srivastava, Nitish, Geoffrey Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: A Simple Way to Prevent Neural Networks from Overfitting. *Journal of Machine Learning Research 15: 1929–58.*

Suter, J., Ebling, S., Volk, M. 2016. Rule-based automatic text simplification for German. *KONVENS 2016*. Bochum, germany, 2016.

Vaswani, Ashish, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *In Advances in Neural Information Pro-cessing Systems.* ArXiv:1706.03762. (pp. 6000-6010).

Zhao, B. and Vogel, S. 2002. Adaptive parallel sentences mining from web bilingual news collection. *In IEEE Int Conf on Data Mining,* pp. 745–748.

Zoph, Barret and Yuret, Deniz and May, Jonathan and Knight, Kevin 2016. Transfer Learning for Low-Resource Neural Machine Translation. *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing.* Austin, Texas Association for Computational Linguistics.

Wang, T., Chen, P., Rochford, J., Quiang J. 2016. Text Simplification using Neural machine translation. *Proceeding of AAAI-16, 4270-4271.*