INTERSPEECH 2019
September 15–19, 2019, Graz, Austria

Assessing the semantic space bias caused by ASR error propagation and its effect on spoken document summarization

Máté Ákos Tündik[1,2], *Valér Kaszás*[1,3], *György Szaszák*[1,4]

[1]Budapest University of Technology and Economics, Budapest, Hungary
[2]Nokia Solutions and Networks Ltd, Budapest, Hungary
[3]Spicy Analytics Ltd, Budapest, Hungary
[4]Telepathy Labs GmbH, Zürich, Switzerland
{tundik,szaszak}@tmit.bme.hu

Abstract

Ambitions in artificial intelligence involve machine understanding of human language. The state-of-the-art approach for Spoken Language Understanding is using an Automatic Speech Recognizer (ASR) to generate transcripts, which are further processed with text-based tools. ASR yields error prone transcripts, these errors then propagate further into the processing pipeline. Subjective tests show on the other hand, that humans understand quite well ASR closed captions despite the word and punctuation errors. Our goal is to assess and quantify the loss in the semantic space resulting from error propagation and also analyze error propagation into speech summarization as a special use-case. We show, that word errors cause a slight shift in the semantic space, which is fairly below the average semantic distance between the sentences within a document. We also show, that punctuation errors have higher impact on summarization performance, which suggests that proper sentence level tokenization is crucial for this task.
Index Terms: spoken language understanding, speech summarization, meaning, semantic similarity, error propagation

1. Introduction

State-of-the-art spoken document summarization is carried out by using an Automatic Speech Recognizer (ASR) tool to obtain a textual transcript and then applying textual document summarization as the subsequent step. The text summarization module typically performs sentence level tokenization first and then applies other operations in the semantic space to provide a ranking of sentences for extractive summary [1], or to provide a semantic encoding for generating an abstractive summary [2, 3]. The most common way of doing projection into the semantic space is through using word embeddings [4].

When working on transcripts yielded by ASR, two types of errors, sentence level tokenization errors and word recognition errors can make the picture more complex. Both of them affects speech summarization, as these errors propagate further in the processing pipeline.

This means that the very first step, sentence level tokenization is already problematic, as punctuation marks and capitalization are missing from raw ASR output. A straightforward way of counteracting this is to provide either direct segmentation through audio [5] or punctuate ASR output [6, 7, 8]. By using the latter approach, not only acoustic, but also linguistic cues can be leveraged. State-of-the-art punctuation is reported in these works to operate between 70-80% in terms of F1-measure, so there is a residual error even when applying the most advanced punctuation approaches.

Regarding ASR, depending on the task and environmental conditions, word error rates (WER) range between 1-30% in applications relevant for industrial exploitation. For less resource rich languages or languages with some specialty – such as rich morphology, compound words etc. – WER can be much higher than in similar complexity tasks for English, although user experience is usually less disturbed than the difference in WER would suggest. i.e. an English ASR working by the same WER is rated lower than a Finnish [9] or Hungarian one [10] by human end-users.

Indeed, humans perform surprisingly well when reading and interpreting ASR produced closed captions, which eventually get punctuated automatically [10]. Obviously, humans can rely on larger context and other modalities, which makes human error repair mechanism more powerful [11, 12], so that the error repair can also remain unconscious. Persons living with hearing impairment were found to perform even better in spontaneously repairing word and especially punctuation errors, their threshold for consciously noticing such errors is much higher [10].

Semantic space projections – especially word embeddings [13] – have become quite popular in Natural Language Processing and Understanding (NLP and NLU). Although such word vector representations are far from perfect in terms of semantic or syntactic consistency and accuracy, they show amazing capabilities in tasks involving semantic processing of the information such as reasoning and analogy tasks [4]. Using word vectors is also state-of-the-art in spoken document summarization. Our interest in this paper is to assess based on objective measures how much the information gets distorted in the semantic space due to word and/or punctuation errors resulting from using an ASR for speech to text purposes. This aspect has been investigated so far mostly from a subjective point-of-view [14, 10], attempts quantifying the effect of ASR error propagation in the semantic space are rare. In [15], the effect of substitution errors was investigated on sentence embeddings, and some works proposed [16] simulation of ASR errors for such analyses. As producing real ASR transcripts is not complicated once audio is available, we preferred to do so and do not simulate, but rather produce ASR errors so that to avoid the mismatch between a true and a simulated ASR transcript. Beside substitution errors, we prefer to address all kind of errors – deletion, insertion, but also punctuation errors – so that to make our experimental setup as close as possible to real use-case scenarios.

This paper is built up as follows: we have raised the problem, explained our motivation and presented related work in this introductory section. The next sections present the used methodology and datasets to assess semantic similarity on sen-

http://dx.doi.org/10.21437/Interspeech.2019-2154

tence and also document level, using a spoken document summarization approach for the latter. Thereafter we demonstrate and discuss our results before drawing our conclusions.

2. Data, ASR and Punctuation

2.1. Scenarios

We are primarily interested in the assessment of semantic bias introduced by the presence of ASR and/or punctuation errors. We create therefore 4 kinds of transcripts to be compared:

MT-MP: Manual Transcript with Manual Punctuation: this is a simple human made gold transcript, which includes punctuation for {. , ? !};

AT-MP: ASR Transcript with Manual Punctuation: we use an ASR transcript, but restore punctuation from the gold transcripts (based on their timestamps, followed by human check);

MT-AP: Manual Transcript with Automatic Punctuation: we remove punctuation from the gold transcript, and predict punctuation automatically;

AT-AP: ASR Transcript with Automatic Punctuation: ASR transcripts are punctuated with the model described in [10].

2.2. Datasets

We perform experiments on English and Hungarian. For **Hungarian**, we use 10 snippets (blocks) from a broadcast news database covering sports news, weather forecasts and news. We have overall 500 sentences and 8k word tokens in total. We use the Kaldi version of the ASR in [17] (with Kaldi decoder) by 6.8%, 10.1%, and 21.4% Word Error Rates (WER) on weather forecasts, broadcast news and sport news, respectively. For AP (automatic punctuation) we use the model from [10] and obtain F1-measures in the range of 60-70% on MT (manual transcript) and 45-50% on AT (ASR transcript).

For **English**, we use the IWSLT2011 translation evaluation campaign dataset which comes along with 9 blocks of TED alike spoken documents [18]. We have overall 800 sentences and 12k word tokens. We use the Kaldi TED-LIUM ASR [19] and its corresponding language model with nnet2 setup for AT and obtain an overall WER of 18.7% (please note that we did not optimize the setup to run the ASR in the optimal operating point, but wished to obtain a WER in this range to have a 'sufficient' number of word errors to assess error propagation). AP is performed with the English punctuation model from [10] by F1-measures in the range of 60-70% on MT and 50-55% on AT.

Both language datasets were summarized by 3 human annotators (summaries up to 10-12 sentences, 3 summaries for each block) to help evaluating semantic bias introduced by AT/AP in a spoken document summarization task.

3. Methodology

We propose and evaluate several approaches to quantify semantic bias, which follow two basic considerations: (i) we calculate semantic similarity between sentence pairs based on word embeddings, while (ii) analyzing the interplay of transcription and punctuation errors is possible through an automatic summarization task. Comparison can hence be sentencewise or documentwise (i.e. in an automatic summarization task).

3.1. Sentencewise semantic similarity

As a first step, the sentence vector representations need to be calculated from word vectors in the sentence. We use GloVe [20] and word2vec [13] for English and Makrai's word vectors for Hungarian [21]. We considered using contextual embeddings and character N-gram augmented word vectors, but we dropped this idea as contextual embeddings are not available for Hungarian at the moment, and adding character N-grams is also counterproductive for Hungarian due to the very rich morphology (word vectors learn morphosyntax and lose from semantic consistency).

Also, using sentence level encoders [22, 23] falls out of our scope in the first step presented in this paper, especially as simpler approaches yet show comparable performance to these heavy and complex approaches [24]. In this way we also overcome difficulties such as missing adaptation for Hungarian, and benefit from all advantages of lightweight unsupervised approaches.We use the following vector representation forms for sentencewise evaluation:

Bag-of-Words (BOW): in the easiest sentence embedding implementation, the word vectors of a sentence are averaged. Optionally stop words can be removed, as they contribute weakly to semantics, we use the python NLTK library for this purpose.

Smooth Inverse Frequency (SIF): SIF sentence embeddings [25], take a weighted average of the word vectors instead of a simple sum (or average). The weight W is calculated as follows:

$$W(w_i) = \frac{a}{a + p(w)}, \tag{1}$$

where a is a smoothing parameter ($a = 0.001$ by default) and $p(w_i)$ is the relative frequency of word w_i in some reference corpus. Commonly used words are hence de-emphasized to favour semantically more relevant words in this formula. Thereafter, in the 'common component removal' step, all SIF vectors in a dataset are concatenated into a matrix. Singular Value Decomposition (SVD) is performed on this matrix, the projections of the SIF sentence embeddings on their first principal component are subtracted from each weighted average, minimizing the impact of semantically irrelevant ('out-of-context') words in this way.

Unsupervised Smoothed Inverse Frequency (uSIF): uSIF [24] follows the SIF approach, but here a is directly computed based on a frequency dictionary, which relieves the need of fine-tuning it. Additionally, the first m principal components, each weighted by the factor $\lambda_1 ... \lambda_m$ are subtracted after SVD:

$$\lambda_i = \frac{\sigma_i^2}{\sum_{j=1}^{m} \sigma_i^2}, \tag{2}$$

where σ_i is the i-th singular value of the sentence embedding matrix. When $m = 1$, uSIF is equivalent to SIF with optimized a. The recommended value for m is 5.

To compute sentence similarity, we perform a pairwise comparison of aligned sentences based on cosine similarity:

$$sim(a, b) = \frac{\sum_{i=0}^{S-1} a_i b_i}{\sum_{i=0}^{S-1} a_i^2 \sum_{i=0}^{S-1} b_i^2} \tag{3}$$

where a and b are the sentence embedding vectors (obtained as BOW, SIF or uSIF) to be compared in the S dimensional sentence embedding space.

There is an alternative to using sentence vectors directly obtained based on word vectors: **Word Mover's Distance** (WMD) is a popular measure to estimate document similarity. WMD [26] uses word vectors computed based on two documents to quantify the distance between the two sentences with

the minimum (weighted) cumulative cost to 'travel' in semantic space to reach the words of the other document. By WMD calculation, Euclidean distance between word vectors is computed, then an Earth mover's distance [27] solver is applied. WMD is available in the popular Gensim python library as well. When using WMD, we obtain the similarity (Word Mover's Similarity, WMS) simply by:

$$WMS = \frac{1}{1+WMD}. \quad (4)$$

3.2. Documentwise semantic similarity

As explained already, the idea behind investigating a spoken document summarization approach w.r.t. semantic distortion caused by ASR and punctuation errors relies on the consideration that this is a typical use case for ASR, where semantics plays the primary role.

We hence prepare summaries for the MT-MP, AT-MP, MT-AP and AT-AP scenarios and compare them based on the Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric family [28]. ROUGE is one of the most commonly used benchmark in text summarization. We refer the reader to [28] for a detailed overview of ROUGE scoring options and measures. In this work we rely on the following ROUGE metrics:

- ROUGE-1: provides a unigram (wordwise) overlap score (recall) between the summary and the reference(s);
- ROUGE-2: provides a bi-gram (wordpair-wise) overlap score (recall) between the summary and the reference(s);
- ROUGE-L: is related to the longest common subsequence (longest common word chain between summary and references; this measure hence takes into account similarity at the sentence level;
- ROUGE-SU4: assess match based on a skip-bigram plus an N-gram based co-occurrence statistics.

As reference, we use human made summaries prepared by 3 independent annotators based on the MT-MP scenario transcripts. We perform the summarization with Gensim [29] and the BM25 scoring function [30]. Our choice for Gensim is motivated by the fact that although Gensim is not the latest and most advanced known summarization tool, but it is simple, well known and widely used, including as well a number of real world and industrial applications, which allows for a global and solid comparison of ASR error propagation impact on summaries. We prefer to analyze a document summarization task instead of document level embedding for the same reasons.

4. Results and Discussion

For the sentencewise evaluation, we compare MT-MP to AT-MP, as aligning sentences between the manual and the automatic punctuation would not be trivial: punctuation errors may alter sentence boundaries, hence the comparison between MP and AP is better suited in the documentwise approach.

Fig. 1 shows sentencewise similarity (BOW, SIF, uSIF and WMS) between MT-MP and AT-MP scenarios in Hungarian (a) and English (b).WER on the x axis here is meant on the sentence level per se. Considering realistic ASR use-cases with WER lower than 30% in Hungarian and than 20% for English[1], the impact on the semantic space is limited to getting still a fair similarity as of 0.8. It is worth to take a look at variances,

[1]Rich morphology Hungarian accounts for higher WER by the similar perceived ASR quality [31]

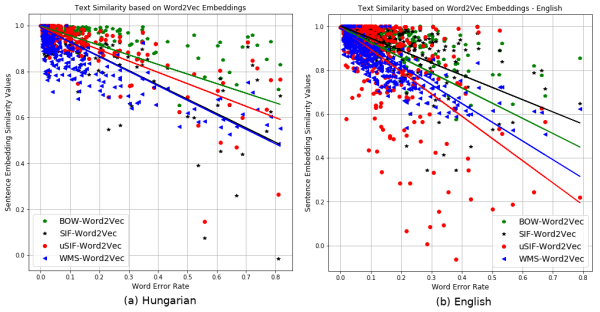

Figure 1: *Sentencewise similarity as a function of word error rate (WER) between MT (gold) and AT (ASR) scenarios.*

Table 1: *Main predefined styles in Word*

Language	**BOW**	SIF	uSIF	WMS
Hungarian	0.97	0.95	0.96	0.92
English	0.94	0.96	0.91	0.90

spreading becomes notable above 20% WER. The means of the overall sentencewise similarities between MT and AT scenarios are shown in Table 1, where we observe very high semantic match despite the word errors.

We run the experiments with both 300-dimensional word2vec and 152-dimensional GloVe word vectors for Hungarian. Using the two types of embeddings we got consistent trends with SIF, uSIF and WMS, hence we plotted only the word2vec results. As expected, there is no significant difference between these two types of word vectors. (Regarding the BOW approach, the two types of word vectors become quasi equivalent when we apply stop word removal for sentence vectors computed as the mean of GloVe word vectors in the sentence. This is understandable as in word2vec stopwords are subsampled [4], whereas they are preserved during GloVe training.)

In order to obtain a kind of benchmark w.r.t. the similarity values seen in Fig. 1 between MT and AT scenarios (with MP), we use the distributions of adjacent sentence similarities within MT type documents. To obtain such histograms, we pair up neighbouring (adjacent) sentences and measure the similarity between them. The idea behind this step is to provide a benchmark so that we can compare the sentencewise semantic bias resulting from word errors to the semantic similarity that can be observed between the sentences of a document. Results in Fig. 2 show uSIF and WMS distributions. In each plot, two histograms are shown: similarity distributions between MT and AT scenarios and similarity distributions between adjacent sentences within MT. The two distributions overlap barely for Hungarian (Fig. 2 a) and English (Fig. 2 b). This means that the semantic bias resulting from word errors is not as high as to make a misrecognized sentence closer to adjacent sentences' meaning than the original meaning. Taking into account that adjacent sentences are typically closer in the semantic space than non adjacent sentences within the same document, this is a rather satisfactory result, which explains why meaning extraction can be still fairly robust from sentences which contain word errors.

Switching to the speech summarization task, we present ROUGE results for all the 4 scenarios in Fig. 3 for Hungarian (a) and English (b). As we have 3 different genres in Hungarian, we could drop down the results based on genres, and we

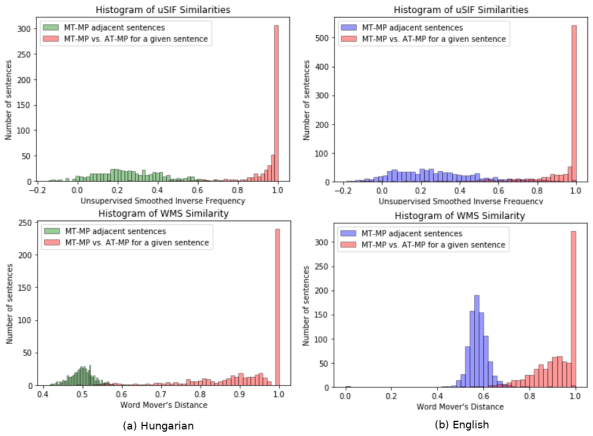

Figure 2: *Histograms showing distribution of semantic similarities (uSIF and WMS) in immediate sentence context versus between MT and AT.*

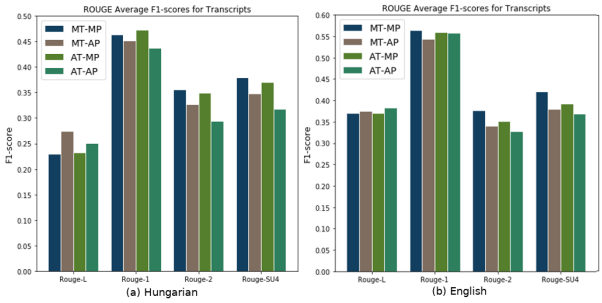

Figure 3: *Evaluation of spoken document summarization on Hungarian and English.*

also show results on individual blocks for Hungarian evaluation data in Fig. 4.

The most relevant cases to compare are the MT-MP (gold) and the real use-case AT-AP, where both transcription and punctuation are derived automatically. Taking a look at the results for the 3 genres in Fig. 4 for Hungarian summarization, we observe that summarization on the MT-MP transcripts is highly correlated with ASR accuracies, most likely because the language model in the ASR and the semantic ranking module in the summary module have to face a task of similar linguistic complexity. Weather forecasts are an exception for this, we suppose that the term frequency based summarization approach is less suitable for this kind of documents. Summarization works the best for broadcast news, despite that weather forecasts can be transcribed by ASR more accurately. Punctuation on weather forecasts is however less accurate [10], the least accurate AP being associated with the sport news [10].

This draws a picture that punctuation errors are more crucial for summarization. This also correlates well with the sentencewise results: word errors caused limited bias in the semantic space at the sentence level, provided that the true sentence level is known (AT-MP). Taking a look at other ROUGE scores in Fig. 3, we also note in ROUGE-2 and ROUGE-SU4 scores that MT-AP is ranked lower than AT-MP (although not with ROUGE-L), and that the gap is larger when switching from MP to AP than when switching from MT to AT. The difference between AT-MP and AT-AP in terms of ROUGE-2 and

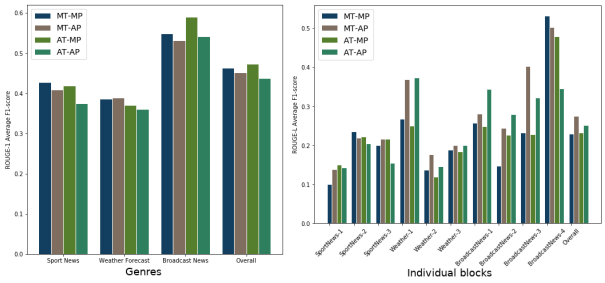

Figure 4: *Spoken document summarization on Hungarian, in genre-wise and block-wise setups.*

ROUGE-SU4 is noticeable. In the AT-AP scenario, word errors already propagate into the AP phase as well. This also confirms that sentence level tokenization errors (that is, punctuation errors) influence summarization at a larger scale than do the word errors. This also suggests that using ASR error robust acoustic (prosodic) cues for punctuation [32] in parallel to the word level models or directly for sentence level tokenization [5] should be considered.

ROUGE-L gets even higher with AP (consistently when going from MT-MP to MT-AP and when going from AT-MP to AT-AP). Since ROUGE-L reflects the longest common subsequence as we have 3 reference summaries for each block, we explain this at the first glance unexpected result by the fact that as reference we match 3 different summaries and hence we have a higher chance to meet this requirement.

5. Conclusions

In this paper we were investigating the semantic bias induced by ASR and punctuation errors. ASR errors already propagate into the automatic punctuation phase in the AT-AP scenario, and the two types of errors propagate further into the speech summarization task.

We demonstrated with a simple similarity assessment framework at the sentence level that word error propagation accounts for less bias in semantic similarity than the semantic difference between two sentences in each other's context. Indeed, similarity distributions for the two cases showed a very weak overlap which suggests that word errors rarely cause dramatic shift in the semantic space at the sentence level (and hence at higher levels, i.e. the document level).

Nevertheless, the true sentence level is lost in ASR, AP has to be applied. Assessing the semantic bias from a speech summarization point of view made us possible to investigate the effect of punctuation errors. We found that punctuation errors account for higher relative difference in ROUGE-2 and ROUGE-SU4 scores than do word errors, although word errors already propagate into the punctuation phase. Analysing summarization on MT-AP showed however, that the current bottleneck in the ASR pipeline arises primarily from a sentence level (punctuation) mismatch rather than from word errors even by WER close to 20%.

6. Acknowledgements

Training was performed on GPUs kindly provided by NVIDIA. Research was supported by the Hungarian National Research, Development and Innovation Office (NKFIH, contract ID: *FK-124413*).

7. References

[1] A. Celikyilmaz and D. Hakkani-Tür, "Discovery of topically coherent sentences for extractive summarization," in *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies-Volume 1*. Association for Computational Linguistics, 2011, pp. 491–499.

[2] P.-E. Genest and G. Lapalme, "Fully abstractive approach to guided summarization," in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, vol. 2, 2012, pp. 354–358.

[3] R. Paulus, C. Xiong, and R. Socher, "A deep reinforced model for abstractive summarization," *arXiv preprint arXiv:1705.04304*, 2017.

[4] T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean, "Distributed representations of words and phrases and their compositionality," in *Advances in Neural Information Processing Systems 26*, C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Weinberger, Eds. Curran Associates, Inc., 2013, pp. 3111–3119. [Online]. Available: http://papers.nips.cc/paper/5021-distributed-representations-of-words-and-phrases-and-their-compositionality.pdf

[5] A. Beke and G. Szaszák, "Automatic summarization of highly spontaneous speech," in *International Conference on Speech and Computer*. Springer, 2016, pp. 140–147.

[6] O. Klejch, P. Bell, and S. Renals, "Sequence-to-sequence models for punctuated transcription combining lexical and acoustic features," in *Acoustics, Speech and Signal Processing (ICASSP), 2017 IEEE International Conference on*. IEEE, 2017, pp. 5700–5704.

[7] A. Öktem, M. Farrús, and L. Wanner, "Attentional parallel RNNs for generating punctuation in transcribed speech," in *International Conference on Statistical Language and Speech Processing*. Springer, 2017, pp. 131–142.

[8] M. A. Tündik and G. Szaszák, "Joint Word-and Character-level Embedding CNN-RNN Models for Punctuation Restoration," in *Cognitive Infocommunications (CogInfoCom), 2018 9th IEEE International Conference on*. IEEE, 2018.

[9] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar, "Unlimited vocabulary speech recognition for agglutinative languages," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*. Association for Computational Linguistics, 2006, pp. 487–494.

[10] M. A. Tündik, G. Szaszák, G. Gosztolya, and A. Beke, "User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning," in *Proc. Interspeech 2018*, 2018, pp. 2628–2632.

[11] A. Postma, "Detection of errors during speech production: A review of speech monitoring models," *Cognition*, vol. 77, no. 2, pp. 97–132, 2000.

[12] B. J. Kröger, E. Crawford, T. Bekolay, and C. Eliasmith, "Modeling interactions between speech production and perception: speech error detection at semantic and phonological levels and the inner speech loop," *Frontiers in Computational Neuroscience*, vol. 10, p. 51, 2016.

[13] T. Mikolov, K. Chen, G. Corrado, and J. Dean, "Efficient estimation of word representations in vector space," in *Proceedings of Workshop at ICLR*, Y. Bengio and Y. LeCun, Eds., vol. 2013, 01 2013.

[14] S. Kafle and M. Huenerfauth, "Effect of speech recognition errors on text understandability for people who are deaf or hard of hearing," in *Proceedings of the 7th Workshop on Speech and Language Processing for Assistive Technologies (SLPAT)*, 2016, pp. 20–25.

[15] R. Voleti, J. M. Liss, and V. Berisha, "Investigating the effects of word substitution errors on sentence embeddings," *arXiv preprint arXiv:1811.07021*, 2018.

[16] E. Simonnet, S. Ghannay, N. Camelin, and Y. Estève, "Simulating asr errors for training slu systems," in *LREC 2018*, 2018.

[17] Á. Varga, B. Tarján, Z. Tobler, G. Szaszák, T. Fegyó, C. Bordás, and P. Mihajlik, "Automatic close captioning for live hungarian television broadcast speech: A fast and resource-efficient approach," in *International Conference on Speech and Computer*. Springer, 2015, pp. 105–112.

[18] M. Federico, S. Stüker, L. Bentivogli, M. Paul, M. Cettolo, T. Herrmann, J. Niehues, and G. Moretti, "The IWSLT 2011 evaluation campaign on automatic talk translation," in *International Conference on Language Resources and Evaluation (LREC)*, 2012, pp. 3543–3550.

[19] A. Rousseau, P. Deléglise, and Y. Esteve, "Ted-lium: an automatic speech recognition dedicated corpus." in *LREC*, 2012, pp. 125–129.

[20] J. Pennington, R. Socher, and C. D. Manning, "Glove: Global vectors for word representation." in *Proceedings of EMNLP*, 2014, pp. 1532–1543.

[21] M. Makrai, "Filtering Wiktionary triangles by linear mapping between distributed models," in *Proceedings of LREC*, 2016, pp. 2776–2770.

[22] D. Cer, Y. Yang, S.-y. Kong, N. Hua, N. Limtiaco, R. S. John, N. Constant, M. Guajardo-Cespedes, S. Yuan, C. Tar *et al.*, "Universal sentence encoder," *arXiv preprint arXiv:1803.11175*, 2018.

[23] A. Conneau, D. Kiela, H. Schwenk, L. Barrault, and A. Bordes, "Supervised learning of universal sentence representations from natural language inference data," *arXiv preprint arXiv:1705.02364*, 2017.

[24] K. Ethayarajh, "Unsupervised random walk sentence embeddings: A strong but simple baseline," in *Proceedings of The Third Workshop on Representation Learning for NLP*, 2018, pp. 91–100.

[25] S. Arora, Y. Liang, and T. Ma, "A simple but tough-to-beat baseline for sentence embeddings," in *International Conference on Learning Representations*, 2016.

[26] M. Kusner, Y. Sun, N. Kolkin, and K. Weinberger, "From word embeddings to document distances," in *International Conference on Machine Learning*, 2015, pp. 957–966.

[27] Y. Rubner, C. Tomasi, and L. J. Guibas, "A metric for distributions with applications to image databases," in *Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271)*. IEEE, 1998, pp. 59–66.

[28] C.-Y. Lin, "Rouge: A package for automatic evaluation of summaries," *Text Summarization Branches Out*, 2004.

[29] R. Mihalcea and P. Tarau, "Textrank: Bringing order into text," in *Proceedings of the 2004 conference on empirical methods in natural language processing*, 2004.

[30] F. Barrios, F. López, L. Argerich, and R. Wachenchauzer, "Variations of the similarity function of textrank for automated summarization," *arXiv preprint arXiv:1602.03606*, 2016.

[31] M. Kurimo, A. Puurula, E. Arisoy, V. Siivola, T. Hirsimäki, J. Pylkkönen, T. Alumäe, and M. Saraclar, "Unlimited vocabulary speech recognition for agglutinative languages," in *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics -*. Morristown, NJ, USA: Association for Computational Linguistics, 2006, pp. 487–494.

[32] A. Moró and G. Szaszák, "A phonological phrase sequence modelling approach for resource efficient and robust real-time punctuation recovery," in *Proceedings of Interspeech*, 2017.