Periodica Polytechnica Electrical Engineering and Computer Science

On the Effects of Automatic Transcription and Segmentation Errors in Hungarian Spoken Language Processing

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

[1] Department of Telecommunication and Media Informatics, Faculty of Electrical Engineering and Informatics, Budapest University of Technology and Economics, H-1117 Budapest, Magyar tudósok körútja 2., Hungary
[2] Nokia Solutions and Networks Ltd., 1083 Budapest, Bókay János u. 36-42, Hungary
[*] Corresponding author, e-mail: tundik@tmit.bme.hu

Received: 21 March 2019, Accepted: 14 April 2019, Published online: 13 June 2019

Abstract

Emerging Artificial Intelligence (AI) technology has brought machines to reach an equal or even superior level compared to human capabilities in several fields; nevertheless, among many other fields, making a computer able to understand human language still remains a challenge. When dealing with speech understanding, Automatic Speech Recognition (ASR) is used to generate transcripts, which are processed with text-based tools targeting Spoken Language Understanding (SLU). Depending on the ASR quality (which further depends on speech quality, the complexity of the topic, environment etc.), transcripts contain errors, which 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. Through word embedding based semantic parsing, the present paper is interested in quantifying the semantic bias introduced by ASR error propagation. As a special use case, speech summarization is also evaluated with regard to ASR error propagation. We show, that despite the higher word error rates seen with the highly inflectional Hungarian, the semantic space suffers least impact than the difference in Word Error Rate would suggest.

Keywords

automatic punctuation, word embedding, semantic similarity, automatic summarization, speech recognition

1 Introduction

With the evolution of Automatic Speech Recognition (ASR), the research studies tend to put more focus on the complex processing of spoken language. In the area of Spoken Language Processing / Understanding (SLP/SLU), the machines are often aimed at human intent inference, and trained for different tasks such as slot filling, keyword spotting, or summarization [1-3].

Two different methods can be applied by the processing of spoken documents; the direct acoustic analysis of the speech stream, or the post-processing of the automatic transcripts. In the latter case, after the speech-to-text transformation, the data can be processed in a textual parser pipeline [4], including stemming, Part-of-Speech tagging, dependency parsing, word sense disambiguation, etc. However, these modules assume that they receive an error-free text on their input, but the ASR-based transcripts can contain different word errors; insertions, substitutions, or deletions. Szaszák et al. [5] showed, that despite of the ASR-error propagation, the syntactic parsing for Hungarian language can be effective through a text-based automatic document summarization approach, by relying on nouns. They also attempted to exploit acoustic features to increase robustness and provide sentence level tokenization based on prosody for a subsequent text-based automatic document summarization approach.

Usually, the lack of segmentation/punctuation of speech or text can be the main bottleneck for the aforementioned modules. Nowadays, the Recurrent Neural Network (RNN)-based approaches offer the most effective solution for automatic punctuation of written language [6, 7]. For Hungarian, we proposed RNN-based punctuation restoration approaches, even with multiple features (character, word, prosody) [8], and the word-level model was evaluated by the end-users as well [9]. Our experience shows that humans can quite well understand error-prone transcripts. Obviously, human error repair mechanisms help in restoring the syntactically and semantically coherent structure, but from a machine-based SLU point

Cite this article as: Tündik, M. Á., Kaszás, V., Szaszák, G. "On the Effects of Automatic Transcription and Segmentation Errors in Hungarian Spoken Language Processing", Periodica Polytechnica Electrical Engineering and Computer Science, 2019. https://doi.org/10.3311/PPee.14052

of view, it is a relevant question to what extent syntax and semantics are affected by ASR and punctuation errors. In this paper we present a separate and a combined analysis of the transcription and punctuation errors through an automatic summarization task.

The word embeddings are popular nowadays [10, 11], because they reflect on the semantic relationships of the words in a vector space model. In our paper, we also investigate how ASR errors propagate into the semantic textual similarity, where we use pre-trained embeddings to represent the sentences of the transcripts of broadcast data. The questions are, how the similarity values react on the changes of Word Error Rate (WER), with a special attention paid to the morphologically rich Hungarian language.

2 Experimental Data

A part of the Hungarian Broadcast Dataset was used for our experiments, which is derived from public broadcasts of the Media Service Support and Asset Management Fund (MTVA), Hungary. The raw data contains reference and ASR-produced [12] transcripts of various TV genres. We selected the transcripts of 10 broadcast blocks with overall 500 utterances of 8143 word tokens in total, among weather forecasts, broadcast news and sport news, because these TV genres were the top most three groups regarding ASR performance, by 6.8 %, 10.1 %, and 21.4 % WER values respectively.

In order to ensure sentence level segmentation for the data an automatic punctuation algorithm was used [13].

Considering manual and machine transcripts and manual and machine punctuation, we created four different types of transcript:

1. Manual transcripts - manual punctuation (MT-MP)
2. ASR transcripts – manual punctuation (AT-MP)
3. Manual transcripts – automatic punctuation (MT-AP)
4. ASR transcripts - automatic punctuation (AT-AP)

The automatic punctuation of these transcripts was done with a word level sequence-to-sequence Recurrent Neural Network (RNN) model. The punctuation marks covered include commas, periods, question marks and exclamation marks. Eventual colons and semicolons were all mapped to commas, whereas all other punctuation marks were simply removed. The details of the model can be found in [13].

For a better understanding of the goals of the proposed experiments, we calculated WER between the sentence pairs of the manually punctuated manual and ASR transcripts. Fig. 1 shows the TV genre-grouped boxplots of WER-values. Fig. 2 illustrates the number of sentence level tokens in the individual blocks. Switching to automatic punctuation (AP) has obviously some impact on the number of sentence tokens.

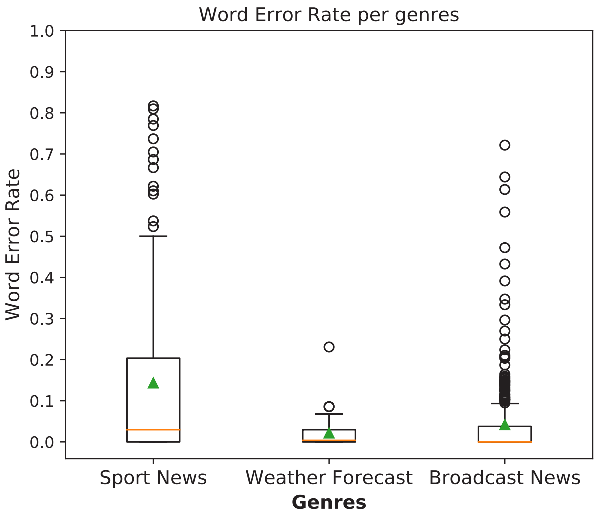

Fig. 1 Word Error Rate values per TV Genres

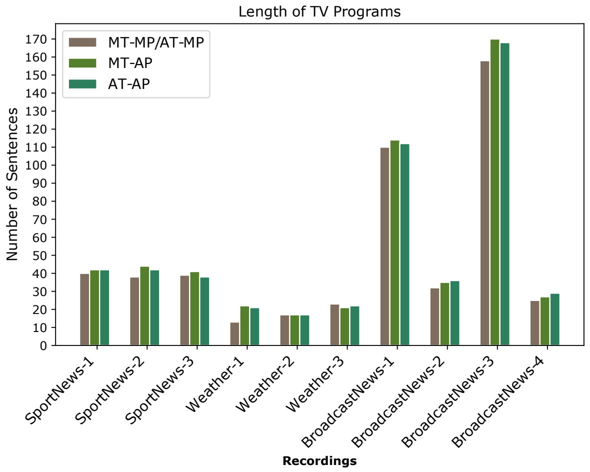

Fig. 2 The number of sentences per recordings, showing the effect of automatic punctuation

It is obvious that the number of sentences is equal for MT-MP and AT-MP, however, the sentence boundaries can be different for the cases MT-AP and AT-AP. Typically, the substitution of comma and period, or the insertion of extra periods increases the number of sentences in these transcript types.

3 Methods

We are interested in the assessment of semantic bias introduced by the presence of ASR and/or punctuation errors. We propose and evaluate several approaches to quantify semantic bias, which follow 2 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. We present these methods in the next sections.

3.1 Measuring Semantic Textual Similarity (STS)

As a first step, the sentence embeddings are calculated from pre-trained word embeddings. To compare two different approaches, we used 152-dimensional GloVe [10] pre-trained embeddings and 300-dimensional word2vec [11] word embeddings as well, provided by Makrai [14] (so we do not perform analysis for character n-gram boosted FastTrack word embeddings [15]).

We used the following methods for sentence embedding determination:

- **Simple Bag-of-Words (BOW):** A common sentence embedding implementation is when all the word vectors of a sentence are averaged to a new vector, also called Bag-of-Word approach. BOW serves as a basis for the next two computation methods.
- **Smooth Inverse Frequency (SIF):** Arora et al. [16] proposed SIF embeddings, which takes the weighted average of the word vectors, where the weight is the $\frac{a}{a+p(w)}$ ratio. In this multiplier, ***a*** is a smoothing parameter (0.001 is offered as a default value) and ***p(w)*** is the word occurrence in a given corpus. Like *Term Frequency*-Inverse Document Frequency (TF-IDF) scheme, commonly used words are de-emphasized to likely merit the semantically relevant content from the rare words. After that, in the "*common component removal*" step, all SIF vectors in a dataset are concatenated into a matrix. Finally, after a Singular Value Decomposition, the projections of the SIF sentence embeddings on their first principal component are subtracted from each weighted average, minimizing the impact of semantically unimportant or "out-of-context" words in this way.
- **Unsupervised Smoothed Inverse Frequency (uSIF):** Proposed by [17], uSIF improves the SIF approach in two ways; ***a*** value is directly computed with the help of the frequency dictionary hence it does not require fine-tuning. Additionally, the first *m* principal components, each weighted by the factor $\lambda_1 \ldots \lambda_m$ are subtracted in the *piecewise common component removal* step. Here

$$\lambda_i = \frac{\sigma_i^2}{\sum_{i=1}^{m} \sigma_i^2} \qquad (1)$$

where σ_i is the i-th singular value of the uSIF sentence embedding matrix. When $m = 1$, it is equivalent the removal step in SIF (In uSIF, *m* optimized empirically ($m = 5$)).
- **Word mover's distance (WMD):** is a popular alternative to estimate document similarity. WMD [18] uses the word embeddings of the words in 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 are computed, then an Earth mover's distance [19] solver is applied. WMD is available in the popular Gensim library[1].

The sentence embeddings are compared with two metrics. For simple BOW-, SIF-, and uSIF-generated vectors we use cosine similarity:

$$\mathrm{sim}\left(d_j, \mathrm{d}_k\right) = \frac{\vec{d}_j \cdot \vec{d}_k}{\left|\vec{d}_j\right|\left|\vec{d}_k\right|} = \frac{\sum w_{\mathrm{i,j}} w_{\mathrm{i,k}}}{\sqrt{\sum w_{\mathrm{i,j}}^2}\sqrt{\sum w_{\mathrm{i,k}}^2}}, \qquad (2)$$

while **Word Mover's Similarity** (WMS) is computed in the following way from WMD:

$$WMS = \frac{1}{1 + WMD}. \qquad (3)$$

Additionally, averages can be computed after filtering stop words, which contain little semantic content (e.g. "is", "the", etc.). We used the Hungarian stop word list of NLTK [20] for this purpose in case of BOW, referred as **BOW-no-sw**. Moreover, SIF and uSIF require word frequency values; we used the frequency dictionary of the Hungarian Webcorpus, with the first 100 000 most frequent words [21, 22].

We are aware of that nowadays the Deep Neural Network-based sentence encoders represent the state-of-the-art methods in this topic, but both the Universal Sentence Encoder [23] by Google, and Infersent [24] by Facebook are language-specific (pre-trained with English data), and adaptation to Hungarian is not possible. Moreover, deep contextualized word representation models (shortly named as Elmo [25]) are also popular, but we prefer simple similarity values for this paper.

[1] https://github.com/RaRe-Technologies/gensim

3.2 Measuring document similarities with ROUGE

The Recall-Oriented Understudy for Gisting Evaluation (ROUGE) metric family [26] is commonly used in text summarization, that is why we selected it for our document similarity related experiments. The basic idea behind ROUGE is to compare word overlap between the automatically produced summary and the reference summaries, which may be human-produced, or derived from highlights. In our cases, we selected a popular vector space modelling tool, Gensim which uses the BM25 [27] scoring function to provide automatic extractive summarization, of AT-AP, AT-MP, and MT-AP transcripts. The ratio parameter of this summarizer was set to provide summaries from one-quarter of all sentences of the original transcripts after the ranking. These summaries were compared with human-produced references, which were provided by three annotators. First, two types of measures are computed, known in information retrieval recall (RCL) and precision (PRC):

$$RCL = \frac{C\left(overlappingwords\right)}{C\left(totalwords \in referencesummary\right)} \tag{4}$$

$$PRC = \frac{C\left(overlappingwords\right)}{C\left(totalwords \in automaticsummary\right)} \tag{5}$$

where C(.) is the count operator. Recall and precision can be combined into the F-measure, which is a single measure easier to compare, especially as recall mostly tends to decrease as precision increases and vice versa when changing the operating point. F-measure is more resistant for operating point shifts:

$$F1 = \frac{2 * RCL * PRC}{\left(RCL + PRC\right)}. \tag{6}$$

Intuitively, the more the words overlap between the system generated and reference summaries, the higher these scores will be. The ROUGE metric family proposes other, more strict measures to assess summaries: counting recall and precision for a sequence of words – known as N-grams – is a common practice, where N is set usually between 1..3, with N = 1 taking us to the single word case presented above. With N-gram metrics, we obtain a stricter, but more accurate evaluation in terms of coverage between the two summaries, as not only the word composition, but also the word order is taken into account, as the word order has high impact on meaning of the complete sentence (or summary). We selected four ROUGE-score variants of F1 by our evaluation:

1. ROUGE-1: Unigram-based score of ROUGE,
2. ROUGE-2: Bigram-based score of ROUGE,
3. ROUGE-L: Longest Common Subsequence (among sequence n-grams)-based score of ROUGE,
4. ROUGE-SU4: Skip-Bigram with a maximum skip distance of 4.

These ROUGE-scores were determined with the ROUGE2-toolkit [28].

4 Results and discussion

4.1 STS-related results

With Fig. 3, let us highlight first on some overall observations regarding 300-dimensional *word2vec*-based sentence embedding similarities affected by WER.

As it is reflected by the density in the left corner, the majority of similarities is bounded by WER=20%. The values are likely between 0.8 and 1.0 in this region, but the variance, especially related to WMS needs further explanation. Our assumption is that it can be hardly compensated in WMS, when the erroneous transcript contains only a few mistakes, but they require moving high semantic distances, to turn back to the reference meaning.

The variance is getting drastically increased around WER=40-50 %, reflecting that the meaning of the sentences is heavily altered from this level. Fortunately, such high WER is not considered as a normal operating point for ASR. According to the slope of the linear regression curves, the simple BOW approach yields the most favorable conditions, however, it involves the stop words with

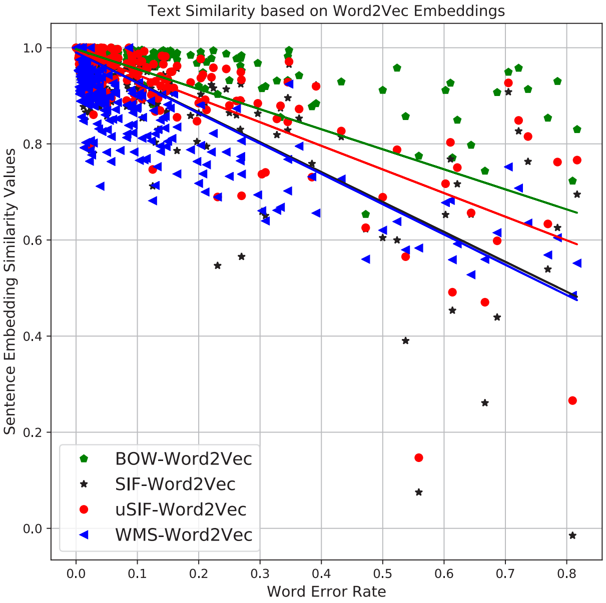

Fig. 3 Word Error Rate - Sentence Embedding value pairs

equal weight, while the SIF and uSIF lines illustrate the effect of more sophisticated, frequency-based word importance in the sentences. Therefore, these latter measures can be considered as reflecting more how the core meaning is affected by ASR word and punctuation errors.

Fig. 4 shows the differences between our *word2vec* and *Glove* embeddings, comparing them through the linear regression-related slopes of each calculated similarities depending on WER.

It can be seen, that in the point of SIF, uSIF and WMS, either choosing *word2vec* or *Glove* shows similar sensitivity on the transcription errors. The main *intra-embedding difference* is pointed out between the BOW approaches, with or without stop words, while *the inter-embedding difference* for BOW is derived from the almost 2:1 ratio of embedding dimensions (300 for *word2vec,* 152 for *Glove*) and 13:1 ratio of embedding vocabulary size (matching for more word both in the MT and AT corpora).

Performing a pairwise comparison on the slopes of the similarity values, between our sensitivity analysis for Hungarian and a similar one for English [29], our results are quite satisfactory. The BOW (regardless of the usage of the stop words), SIF and uSIF are more robust (around rel. 5-10 %) for our morphologically rich language than for English, which also suggests that word-embedding based spoken language processing task incl. automatic summarization may be less affected by ASR errors for Hungarian.

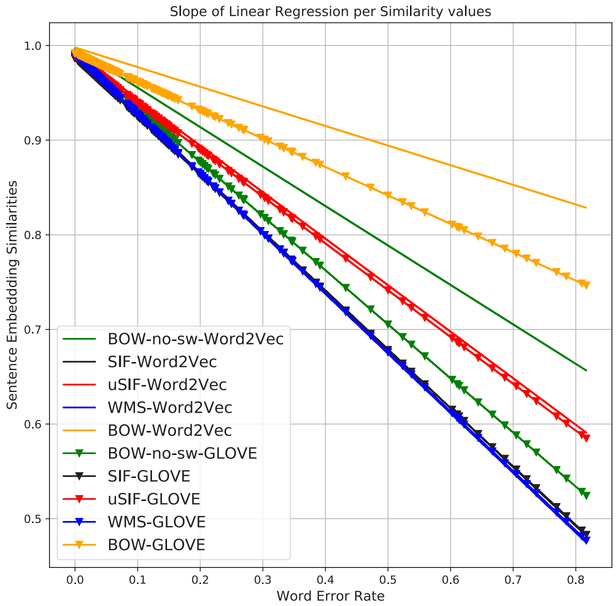

Fig. 4 Word2Vec and Glove comparison regarding transcription error sensitivity

In addition to our overall analysis, we provide some insight into TV genre-based and recording-based differences as well, shown in Figs. 5-7 for the cases of different similarities. As Fig. 5 shows, the average SIF values are similar for Broadcast News and Weather Forecast (0.95), while it is around 0.9 for Sport News – samples, due to higher WER.

Fig. 6 depicts that e.g. in case of BOW-based comparison with cosine similarities by TV genres, the superior performance of Word2Vec is only relative 2-3 % compared to the Glove embeddings. According to Fig. 7, the tendency is similar in the performance by WMS as well.

We summarize our results in Table 1. According to our measurements, the application of uSIF would be advisable. Albeit, the BOW is more robust to ASR-errors, uSIF is able to catch semantically more relevant context than BOW does. Unfortunately, we cannot evaluate these metrics more objectively because Hungarian lacks corpus

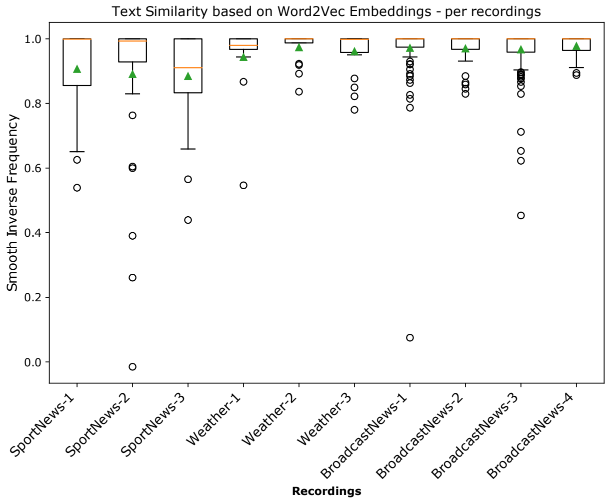

Fig. 5 Smooth Inverse Frequency-analysis per recordings

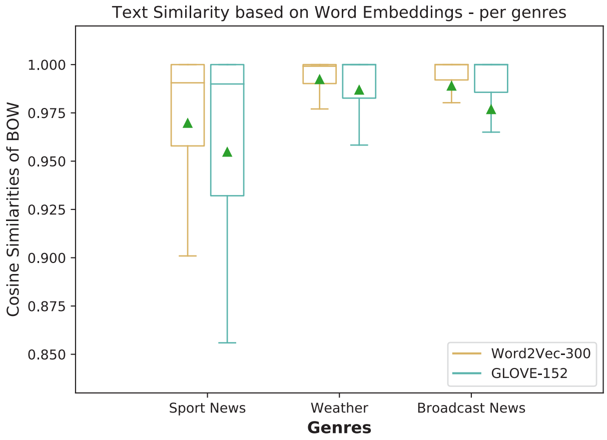

Fig. 6 BOW values with different word embeddings per TV genres

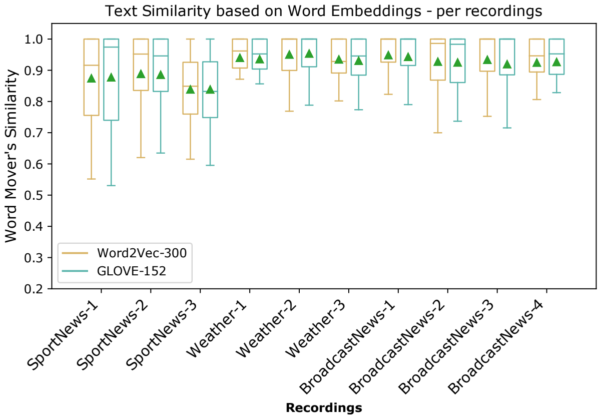

Fig. 7 WMS values with different word embeddings per recordings

Table 1 Results of STS-experiments

	Average Sentence Embedding Similarity Values				
Text Categories	BOW	BOW no-sw	SIF	uSIF	**WMS**
SportNews-W2V	0.970	0.933	0.894	0.920	0.867
SportNews-GL	0.955	0.91	0.897	0.920	0.867
Weather-W2V	0.992	0.982	0.961	0.975	0.941
Weather-GL	0.987	0.974	0.961	0.973	0.939
Broadcast-W2V	0.988	0.982	0.970	0.974	0.937
Broadcast-GL	0.977	0.968	0.963	0.968	0.929
Overall-W2V	0.985	0.971	0.951	0.961	0.921
Overall-GL	0.973	0.955	0.948	0.957	0.915

including semantic similarity scores, but it can be an important task for the near future.

4.2 Document Similarity-related results

First, Fig. 8 shows the overall results of our extractive summarization experiments, evaluated with four ROUGE-scores, for the four types of transcripts.

Please note that usually the F1-scores are well below 100 % for ROUGE, as was proven by [30]. As it is expected, the ROUGE-L shows the lowest values and some anomalies in the scores, according to the strictest criterion for n-gram matching, and we also assume that the short length of the original documents is the second important factor. Of course, as we did it for automatic punctuation [9], a user-focused evaluation would likely allow deeper insight into the phenomenon, as outlined by the authors of [31], who performed this for automatic summarization. On the contrary, the ROUGE-1 provides the highest scores due to unigram approach. Comparing the texts with automatic punctuation, switching from manual to automatic transcripts yields rel. 5-9 % performance drop in overall. Except for ROUGE-L, the manually punctuated transcripts perform better than the automatically segmented variants. However, the rel. 2-3 % differences in performance show the superiority of MT-MP and AT-MP in 2-2 cases. We explain this phenomenon with the relatively low WER for the weather forecast - and broadcast news – related transcripts. Fig. 9 and Fig. 10 provide some details about TV genre-level and recording-level diversities.

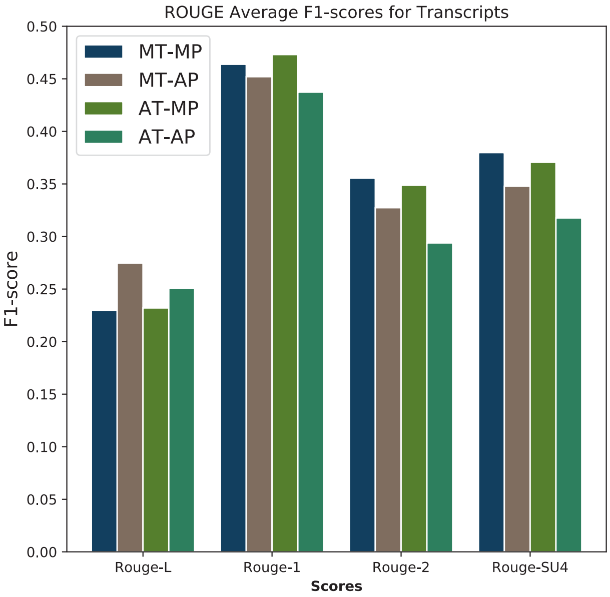

Fig. 8 Summary of Rouge-scores by Transcripts

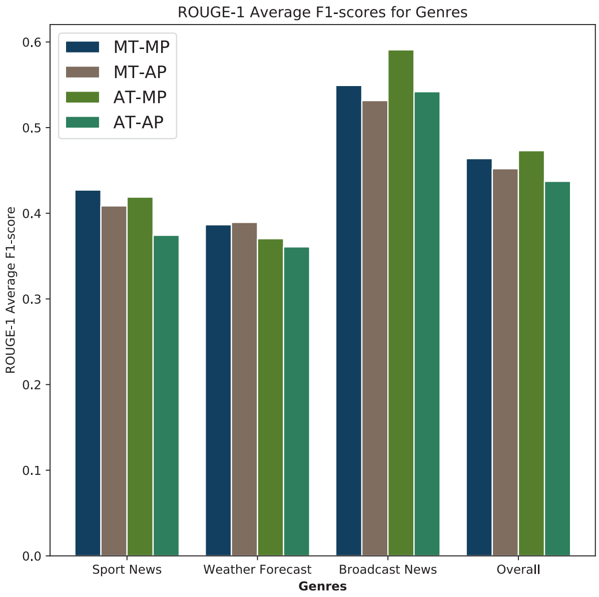

Fig. 9 TV Genre-level Rouge-1 scores

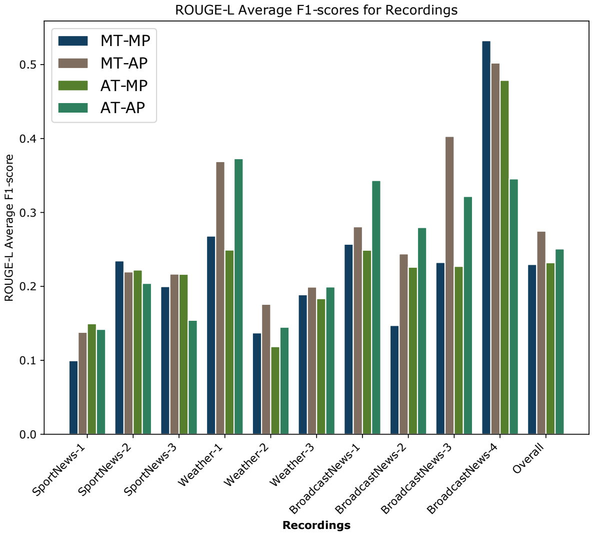

Fig. 10 Recording-level Rouge-L scores

All in all, the results show that the effect of punctuation errors is more significant than the effect of transcription errors on automatic summarization. More precisely, as usually the sentence boundaries count in these tasks, and question and exclamation marks are underrepresented in the investigated topics, in this case, the primary effect is derived from the period-related errors. We confirmed our hypotheses with Pearson-correlations at p=0.05 level. The significant results on p=0.05 level are shown in Table 2.

According to the correlation values of AT-AP category, the interplay of punctuation and transcription errors are highly pronounced. However, we could not measure such significance for MT-AP category, but we also experienced the dominant effect of punctuation errors on Rouge-scores there.

5 Conclusion

In our paper, we investigated the effects of transcription and punctuation errors on SLU-related tasks, such as STS and automatic summarization. We showed that word embeddings are relatively robust to ASR-error propagation in Hungarian, moreover, the automatically punctuated texts yield fairly comparable results to the reference transcripts. According to our results, the capabilities of Hungarian ASR-systems extended with automatic punctuation post-processing module allow for the improvement of automatic summarization system, also involving word embeddings. This points to a possible application in future summarizing systems built or adapted for the Hungarian language, which can be used for both online media content, television, or hearing aid transcription applications.

Table 2 Punctuation Error – Rouge Score Correlations

Transcript Type	Rouge Score Type	Punctuation Score	Pearson Correlation (p=0.05)
AT-AP	L	F1 (overall)	0.785
AT-AP	L	Slot Error Rate	-0.786
AT-AP	1	F1 (overall)	0.674
AT-AP	1	Slot Error Rate	-0.635
AT-AP	2	F1 (overall)	0.723
AT-AP	2	Slot Error Rate	-0.687
AT-AP	SU-4	F1 (overall)	0.713
AT-AP	SU-4	Slot Error Rate	-0.672

Acknowledgement

The authors would like to thank the support of the Hungarian National Research, Development and Innovation Office (NKFIH) under contract ID FK-124413 “Enhancement of deep learning based semantic representations with acoustic-prosodic features for automatic spoken document summarization and retrieval”. The authors would like to thank Balázs Tarján and Péter Mihajlik for their help and assistance in processing the broadcast data and also providing ASR transcripts for the experiments.

References

[1] Beke, A., Szaszák, G. "Automatic Summarization of Highly Spontaneous Speech", In: Ronzhin A., Potapova R., Németh G. (eds.) Speech and Computer. SPECOM 2016. Lecture Notes in Computer Science, vol 9811. Springer, Cham, pp. 140–147. https://doi.org/10.1007/978-3-319-43958-7_16

[2] Ward, W. "The CMU air travel information service: Understanding spontaneous speech", In: HLT ‘90 Proceedings of the workshop on Speech and Natural Language, Hidden Valley, Pennsylvania, 1990, pp. 127–129. https://doi.org/10.3115/116580.116621

[3] Mesnil, G., He, X., Deng, L., Bengio, Y. "Investigation of Recurrent-Neural-Network Architectures and Learning Methods for Spoken Language Understanding", In: INTERSPEECH 2013 14th Annual Conference of the International Speech Communication Association, 2013, pp. 3771–3775.

[4] Váradi, T., Simon, E., Sass, B., Gerocs, M., Mittelholtz, I., Novák, A., Indig, B., Prószéky, G., Vincze, V. "Az e-magyar digitális nyelvfeldolgozó rendszer" (The „e-magyar” digital language processing system), In: XIII. Magyar Számítógépes Nyelvészeti Konferencia (MSZNY2017), Szeged, Hungary, 2017, pp. 49–60. (in Hungarian)

[5] Szaszák, G., Tündik, M. Á., Beke, A. "Summarization of Spontaneous Speech using Automatic Speech Recognition and a Speech Prosody based Tokenizer", In: Proceedings of the 8th International Joint Conference on Knowledge Discovery, Knowledge Engineering and Knowledge Management (IC3K 2016), Porto, Portugal, 2016, pp. 221–227. https://doi.org/10.5220/0006044802210227

[6] Tilk, O., Alumäe, T. "Bidirectional Recurrent Neural Network with Attention Mechanism for Punctuation Restoration", In: Proceedings of the 17th Annual Conference of the International Speech Communication Association (INTERSPEECH 2016), San Francisco, CA, USA, 2016, pp. 3047–3051. https://doi.org/10.21437/Interspeech.2016-1517

[7] Yi, J., Tao, J., Wen, Z., Li, Y. "Distilling Knowledge from an Ensemble of Models for Punctuation Prediction", In: Proceedings of the 18th Annual Conference of the International Speech Communication Association (INTERSPEECH 2017), Stockholm, Sweden, 2017, pp. 2779–2783. https://doi.org/10.21437/Interspeech.2017-1079

[8] Tündik, M. Á., Szaszák, G. "Kombinált központozási megoldások magyar nyelvre pehelysúlyú neurális hálózatokkal" (Combined punctuation approaches for Hungarian with lightweight neural networks), In: XV. Magyar Számítógépes Nyelvészeti Konferencia, Szeged, Hungary, 2019, pp. 275–286.

[9] Tündik, M. Á., Szaszák, G., Gosztolya, G., Beke, A. "User-centric Evaluation of Automatic Punctuation in ASR Closed Captioning", In: Proceedings of the 19th Annual Conference of the International Speech Communication Association (INTERSPEECH 2018), Hyderabad, India, pp. 2628–2632, 2018. https://doi.org/10.21437/Interspeech.2018-1352

[10] Pennington, J., Socher, R.. Manning, C. D. "GloVe: Global Vectors for Word Representation", In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), Doha, Qatar, 2014, pp. 1532–1543. https://doi.org/10.3115/v1/D14-1162

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G., Dean, J. "Distributed representations of words and phrases and their compositionality", In: NIPS'13 Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2, Lake Tahoe, Nevada, USA, 2013, pp. 3111–3119.

[12] Varga, Á., Tarján, B., Tobler, Z., Szaszák, G., Fegyó, T., Bordás, C., Mihajlik, P. "Automatic Close Captioning for Live Hungarian Television Broadcast Speech: A Fast and Resource-Efficient Approach", In: Ronzhin A., Potapova R., Fakotakis N. (eds) Speech and Computer. SPECOM 2015. Lecture Notes in Computer Science, vol 9319. Springer, Cham, 2015, pp. 105–112. https://doi.org/10.1007/978-3-319-23132-7_13

[13] Tündik, M. Á., Tarján, B., Szaszák, G. "Low Latency MaxEnt- and RNN-Based Word Sequence Models for Punctuation Restoration of Closed Caption Data", 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, 2017, pp. 155–166. https://doi.org/10.1007/978-3-319-68456-7_13

[14] Makrai, M. "Filtering Wiktionary triangles by linear mapping between distributed word models", In: Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016), Portorož, Slovenia, 2016, pp. 2766–2770.

[15] Joulin, A., Grave, E., Bojanowski, P., Douze, M., Jégou, H., Mikolov, T. "FastText.zip: Compressing text classification models", 2016. [online] Available at: https://arxiv.org/abs/1612.03651 [Accessed: 20 March 2019]

[16] Arora, S., Liang, Y., Ma, T. "A Simple but Tough-to-Beat Baseline for Sentence Embeddings", In: ICLR 2017: 5th International Conference on Learning Representations, Toulon, France, 2017.

[17] Ethayarajh, K. "Unsupervised Random Walk Sentence Embeddings: A Strong but Simple Baseline", In: Proceedings of the 3rd Workshop on Representation Learning for NLP, Melbourne, Australia, 2018, pp. 91–100.

[18] Kusner, M. J., Sun, Y., Kolkin, N. I., Weinberger, K. Q. "From Word Embeddings to Document Ddistances", In: Proceedings of the 32nd International Conference on Machine Learning, Lille, France, 2015, pp. 957–966.

[19] Rubner, Y., Tomasi, C., Guibas., L. J. "A metric for distributions with applications to image databases", In: Sixth International Conference on Computer Vision (IEEE Cat. No. 98CH36271), Bombay, India, 1998, pp. 59–66. https://doi.org/10.1109/ICCV.1998.710701

[20] Bird, S. "NLTK: the natural language toolkit", In: COLING-ACL '06 Proceedings of the COLING/ACL on Interactive presentation sessions, Sydney, Australia, 2006, pp. 69–72. https://doi.org/10.3115/1225403.1225421

[21] Halácsy, P., Kornai, A., Németh, L., Rung, A., Szakadát, I., Trón, V. "Creating open language resources for Hungarian", In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), Lisbon, Portugal, 2004, pp. 203–210.

[22] Kornai, A, Halácsy, P., Nagy, V., Oravecz, Cs., Trón, V., Varga, D. "Web-based frequency dictionaries for medium density languages", In: Proceedings of the 2nd International Workshop on Web as Corpus, Trento, Italy, 2006, pp. 1–8.

[23] Cer, D., Yang, Y., Kong, S., Hua, N., Limtiaco, N., …, Kurzweil, A. "Universal sentence encoder", arXiv preprint, 2018. [online] Available at: https://arxiv.org/abs/1803.11175 [Accessed: 20 March 2019]

[24] Conneau, A., Kiela, D., Schwenk, H., Barrault, L., Bordes, A. "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data", In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, Copenhagen, Denmark, 2017, pp. 670–680. https://doi.org/10.18653/v1/D17-1070

[25] Peters, M. E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K., Zettlemoye, L. "Deep contextualized word representations", In: Proceedings of NAACL-HLT 2018, New Orleans, Louisiana, USA, pp. 2227–2237. https://doi.org/10.18653/v1/N18-1202

[26] Lin, C.-Y. "Rouge: A package for automatic evaluation of summaries", In: Workshop on Text Summarization Branches Out (WAS 2004), Barcelona, Spain, 2004.

[27] Barrios, F., López, F., Argerich, L., Wachenchauzer, R. "Variations of the similarity function of textrank for automated summarization", 2016. [online] Available at: https://arxiv.org/abs/1602.03606 [Accessed: 20 March 2019]

[28] Ganesan, K. "ROUGE 2.0: Updated and Improved Measures for Evaluation of Summarization Tasks", arXiv preprint, 2018. [online] Available at: https://arXiv:1803.01937 [Accessed: 20 March 2019]

[29] Voleti, R., Liss M., J., Berisha, V. "Investigating the Effects of Word Substitution Errors on Sentence Embeddings", arXiv preprint, 2018. [online] Available at: https://arxiv.org/abs/1811.07021 [Accessed: 20 March 2019]

[30] Schluter, N. "The limits of automatic summarisation according to ROUGE", In: Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers, Valencia, Spain, 2017, pp.41–45.

[31] Feifan, L., Yang, L. "Correlation between rouge and human evaluation of extractive meeting summaries", In: Proceedings of ACL-08: HLT, Short Papers (Companion Volume), Columbus, Ohio, USA, 2008, pp. 201–204.