INTERSPEECH 2017
August 20–24, 2017, Stockholm, Sweden

Evaluating Automatic Topic Segmentation as a Segment Retrieval Task

Abdessalam Bouchekif[1], *Delphine Charlet*[2], *Géraldine Damnati*[2], *Nathalie Camelin*[1], *Yannick Estève*[1]

[1] LIUM, University of Le Mans, France.
[2] Orange Labs, 2 avenue Pierre Marzin 22300, Lannion, France.
[1]firstname.lastname@univ-lemans.fr, [2]firstname.lastname@orange.com

Abstract

Several evaluation metrics have been proposed for topic segmentation. Most of them rely on the paradigm that segmentation is mainly a task that detects boundaries, and thus are oriented on boundary detection evaluation. Nevertheless, this paradigm is not appropriate to get homogeneous chapters, which is one of the major applications of topic segmentation. For instance on Broadcast News, topic segmentation enables users to watch a chapter independently of the others.

We propose to consider segmentation as a task that detects homogeneous segments, and we propose evaluation metrics oriented on segment retrieval. The proposed metrics are experimented on various TV shows from different channels. Results are analysed and discussed, highlighting their relevance.

1. Introduction

In the field of language technology, topic segmentation of a written or spoken document is the task of splitting the document into homogeneous segments, by placing boundaries within the document. Segments can be considered homogeneous according to a variety of dimensions: for instance, in terms of speakers (speaker segmentation), topics (topic segmentation), *etc.* This segmentation is usually a pre-processing step for many other tasks, but it could also have a direct applicative purpose. For example, in topic segmentation of Broadcast News, the aim can be to get homogeneous chapters, so that a user can watch a chapter independently of the others. As the automatic segmentation can be a hard task, it requires appropriate evaluation metrics. Many works have been done to propose evaluation metrics for segmentation. Most of them rely on the paradigm that segmentation is mainly a task that detects boundaries, and thus are oriented on boundary detection evaluation. We propose to consider segmentation as a task that detect homogeneous segments, and we propose evaluation metrics oriented on segment detection.

2. Related work

Recall/Precision are standard evaluation measures for information retrieval tasks, and are often applied to evaluation of topic segmentation. For this, we compare the position of reference and hypothesis segments, with few seconds tolerance (usually $10s$). A reference boundary can be detected (true positive) or missed (false negative), and an hypothesis boundary can match a reference boundary (true positive) or can be a false alarm (false negative). A recall of 100% means that all reference boundaries have been correctly found. A precision of 100% means that all the boundaries proposed by the system are correct.

The measure p_k [1] is based on the principle of a sliding window of size k traveling in parallel on reference and hypothesis segmentations. The principle of p_k is to count the number of times the two ends of the window belong to the same segment, both in the segmentation of reference and in the hypothesis. So, higher score (*e.g* $p_k = 1$) means that the system has a worse quality, otherwise (*e.g* $p_k = 0$) it has a better quality. p_k has several shortcomings [2]. First, some errors are not penalized or under-penalized, this happens when multiple boundaries occur in sliding window. Also, p_k is sensitive to variation of the segment and window size. Moreover, the meaning of the score is not directly intelligible.

The *Windowdiff* (*WD*) [2] metric has been proposed for solving some of the p_k drawbacks. It is also based on a sliding window, but *WD* computes the difference of boundaries number between reference and hypothesis segmentation in sliding window. Even if *WD* overcomes several drawbacks of p_k, it is not perfect. Indeed, the first and last boundaries of a hypothesis are less penalized [3]. The score can be greater than 1, so it can no longer be assimilated to a percentage [4].
In order to go over some of these problems, [5] proposed to normalize *Windowdiff*. In the same way, the authors of [6] proposed *WinPR* which is derived from *WD*, but differs on one main point: *Windowdiff* evaluates boundary positions, while *WinPR* evaluates regions (or windows).
In [7], authors proposed a different approach called *segmentation similarity* that quantifies the similarity between two segmentations as the proportion of boundaries that are not transformed when comparing them using editing distance. Finally, [8] proposed a new series of metrics derived from an adaptation of boundary editing distance [7].

3. Proposed approach

The evaluation metrics described above evaluate the quality of the topic segmentation by computing a classification rate (p_k, *WD* , *WinPR*) or by comparing the position of the topic boundaries of reference and hypothesis (*Recall/Precision*). *Recall/Precision* are more adapted for navigation application [1]. Indeed, these metrics give the rate of correctly returned boundaries and the number of false alarms. For example, a recall of 100% and a precision of 40% means that the user will necessarily find the beginning of each topic of the show or document. However, 60% of the boundaries are false alarms. However, topic segmentation is also used in other tasks such as information retrieval (*e.g* google news), automatic summary, topic modeling. These applications are more interested in the quality of segmentation produced in terms of segments.
We propose two metrics evaluating the quality of a topic segmentation, either by the number or by the duration of the correctly retrieved segments :

- *CovN* relies on the number of segments correctly proposed. The best segmentation corresponds to the one that

[1]The user can access more quickly to the parts he is the most interested in.

http://dx.doi.org/10.21437/Interspeech.2017-1231

proposes the maximum of correct segments.

- *CovD* relies on the duration of correctly proposed segments. The best segmentation corresponds to the one that offers correct segments covering most of the show/document.

This evaluation requires to define what a correct segment is. This definition implies to make a matching between the reference and hypothesis segments.

3.1. Matching between reference and hypothesis segments

For each reference segment, we find the hypothesis segment that covers the most of it.
Given a reference segment R_i, we compute the percentage of duration of R_i which overlaps with $H_j (j = 1, .., k)$, where k is the number of hypothesis segments covered by R_i. This percentage is denotes as $Cov_{R_i \to H_j}$,

Reciprocally, for the hypothesis segment H_i, $Cov_{H_i \to R_j}$ is the percentage of duration of H_i which overlaps with R_j.
In Figure 1, the matching between reference segmentation (R_1, R_2, R_3, R_4) and an hypothesized segmentation (H_1, H_2, H_3, H_4) is illustrated by arrows. For example, R_4 matches H_4, because $Cov_{R_4 \to H_4} > Cov_{R_4 \to H_3} > Cov_{R_4 \to H_2}$.

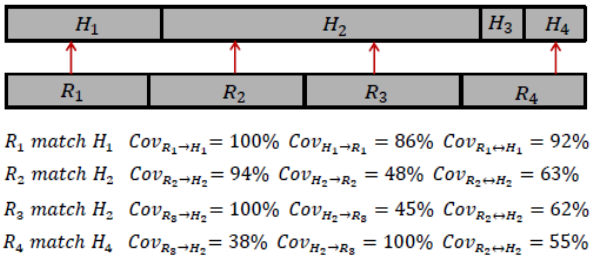

Figure 1: *Example of harmonic coverage computation.*

3.2. What is a correct segment?

The bidirectional coverage between R_i and H_j, denoted $Cov_{R_i \leftrightarrow H_j}$ is defined in eqation 1 as the harmonic mean of $Cov_{R_i H_j}$ and $Cov_{H_j \to R_i}$:

$$Cov_{R_i \leftrightarrow H_j} = \frac{2 \times Cov_{R_i \to H_j} \times Cov_{H_j \to R_i}}{Cov_{R_i \to H_j} + Cov_{H_j \to R_i}}. \quad (1)$$

$Cov_{H_j \to R_i}$ can be seen as the temporal precision of the system's response while $Cov_{R_i \to H_j}$ can be seen as the temporal recall of the reference segment and $Cov_{R_i \leftrightarrow H_j}$ can be seen as *F-measure*.

We consider that the reference segment R_i is correct if $Cov_{R_i \leftrightarrow H_j}$ is greater than a certain threshold γ, else it is considered incorrect. In Fig. 1, if the considering threshold is $\gamma = 85\%$, only the segment H_1 is correct ($Cov_{R_1 \leftrightarrow H_1} > 85\%$).

3.3. CovN and CovD measures

For a given set of N_R reference segments and their associated hypothesis segments, we can compute the number of correctly retrieved reference segments, for a given threshold γ. If we define, for any reference segment R_i, the binary value $\varphi_\gamma(R_i)$, which indicates that the segment is correctly retrieved for the required threshold γ:

$$\varphi_\gamma(R_i) = \begin{cases} 1 & \text{if } Cov_{R_i \leftrightarrow H_j} > \gamma \\ 0 & \text{sinon.} \end{cases}$$

We can then compute the number of reference segments correctly retrieved ($\sum_{i=1}^{N_R} \varphi_\gamma(R_i)$), and consequently, we can compute $\mathcal{R}_\mathcal{N}$ the recall rate of reference segments, as the percentage of correctly retrieved reference segments:

$$\mathcal{R}_N = \frac{1}{N_R} \sum_{i=1}^{N_R} \varphi_\gamma(R_i) \quad (2)$$

Similarly, starting from the hypothesis segment, we can consider that an hypothesis segment is correct if its harmonic coverage is over the threshold γ. We define for any of the N_H hypothesis segments, the binary value $\varphi_\gamma(H_i)$ which indicates if the hypothesis segment is correct considering the threshold γ. Hence, the number of hypothesis segments which are correct is calculated by $\sum_{i=1}^{N_H} \varphi_\gamma(H_i)$. We can then compute the precision rate of hypothesis segments $\mathcal{P}_\mathcal{N}$, as follow:

$$\mathcal{P}_N = \frac{1}{N_H} \sum_{i=1}^{N_H} \varphi_\gamma(H_i) \quad (3)$$

As we get the precision rate $\mathcal{P}_N$ and the recall rate $\mathcal{R}_N$, we can also calculate a single evaluation metric Cov_N.

$$CovN = \frac{2 \times \mathcal{R}_\mathcal{N} \times \mathcal{P}_\mathcal{N}}{\mathcal{R}_\mathcal{N} + \mathcal{P}_\mathcal{N}} \quad (4)$$

The measure $CovD$ can be seen as the version of $CovN$ that takes into account the duration of segments. Indeed, the segments size can range from a few seconds to a few minutes. Generally, the longest segments are the most important. So we give to long segments more weight (*e.g* actuality information) in comparison to the small ones (*e.g* brief information). $CovD$ weights each segment by its duration.

$$\mathcal{R}_D = \frac{1}{\sum_{i=1}^{N_R} d(R_i)} \sum_{i=1}^{N_R} d(R_i) \varphi_\gamma(R_i) \quad (5)$$

$$\mathcal{P}_D = \frac{1}{\sum_{i=1}^{N_H} d(H_i)} \sum_{i=1}^{N_H} d(H_i) \varphi_\gamma(H_i) \quad (6)$$

where $d(S)$ is the duration of segment S. We can then compute their harmonic mean Cov_D.
CovN and *CovD* are complementary. Indeed, the measure *CovN* returns the rate of the correct segments and *CovD* gives additional information like the duration of the segments. For example, considering a TVBN show that lasts 10 minutes which contains 4 segments, if $CovN = 25\%$ and $CovD = 50\%$, we can conclude that the system returns only one correct segment which lasts about half the size of the show.

4. Analysis of the *CovN*

In this part, we study the behaviour of the *CovN* measure according to the type of errors (insertion or deletion). For illustration purpose, we will use Fig. 2 and Fig. 3.
Let R_i be the i^{th} reference segment. We assume that the system detects two segments H_j and H_k covering the segment R_i (false alarm error). We identify several possible cases :

- 1^{st} case: $Cov_{R_i \leftrightarrow H_j} < \gamma$ and $Cov_{R_i \leftrightarrow H_k} < \gamma$. Both segments are considered as false. That is the case, for the segments H_1 and H_2 of Hyp_2.

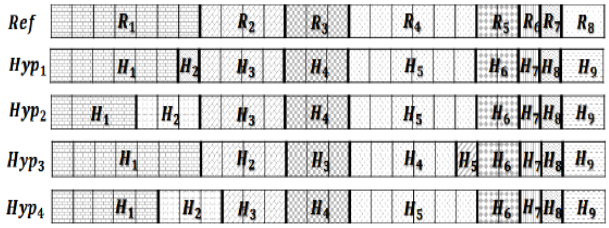

Figure 2: *Penalization of false alarms*

- 2^{nd} case: one of them is correct (the coverage exceeds threshold γ), only one segment is considered as wrong. That is the case, for the segment H_2 of Hyp_1. However, the segment H_1 is considered correct ($Cov_{R_1 \leftrightarrow H_1} > \gamma$).
- 3^{rd} case: insertion of a false segment which produces the edge effect (*i.e* the impact of a false segment on the neighbors segments). That is the case, for the segment H_2 of Hyp_4, $CovN$ considers that H_1, H_2 and H_3 are incorrect.

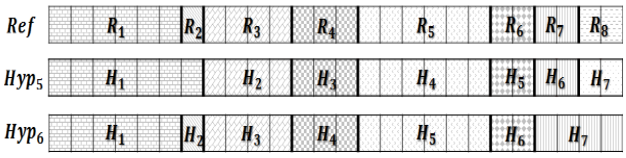

Figure 3: *Penalization of missing errors*

Let's now consider the errors of Fig.3. For the two reference segments R_i and R_j, the system proposes only one segment H_k. In some cases, the measure $CovN$ is less sensitive to non-detected small segments when the neighbors segments are correct.

We compute $CovN$ for each segmentation of Fig.2 and Fig.3, with $\gamma = 85\%$:

	Segments			Boundaries		
Hyp	$\mathcal{R}_N$	$\mathcal{P}_N$	$CovN$	R	P	Fm
Hyp_1	$\frac{8}{8}$	$\frac{8}{9}$	94.1%	$\frac{7}{7}$	$\frac{7}{8}$	93.3%
Hyp_2	$\frac{7}{8}$	$\frac{7}{9}$	82.5%	$\frac{7}{7}$	$\frac{7}{8}$	93.3%
Hyp_3	$\frac{7}{8}$	$\frac{7}{9}$	82.5%	$\frac{7}{7}$	$\frac{7}{8}$	93.3%
Hyp_4	$\frac{6}{8}$	$\frac{6}{9}$	70.6%	$\frac{6}{7}$	$\frac{6}{8}$	80.0%
Hyp_5	$\frac{7}{8}$	$\frac{7}{7}$	93.3%	$\frac{6}{7}$	$\frac{6}{6}$	92.3%
Hyp_6	$\frac{6}{8}$	$\frac{6}{7}$	80.0%	$\frac{6}{7}$	$\frac{6}{6}$	92.3%

The analysis of the *CovN* behaviour allowed us to understand that:

- The deletion and insertion errors do not go unnoticed (even with less severe cases).
- A lack of a small part of reference segment (insertion in the beginning or end of a false segment of small size) is less penalized than missing of false segment of large size or in the middle of reference segment. Similarly, adding a small portion in the beginning or in the end of the reference segment is less penalized than adding in the middle or if the false portion is a large size.
- Over-segmentation is slightly more penalized than sub-segmentation, if the error has no impact on neighbors. In the opposite case, $CovN$ takes a little higher value with an insertion.

Let notice that in some applications such as information retrieval, it is better to add a small segment rather than remove it.

One of the advantages of this metric is that it makes possible to analyse the links between the performance of retrieval for each segment, and some properties of these segments (*e.g.* length, linguistic content, ...), and could contribute to a better understanding of the factors that help or hamper the segmentation process. In [9], $CovN$ metric has been used to evaluate a topic titling system[2]. Obviously, topic titling is applicable only to correct segments, so $CovN$ is considered as one source in titling metric. This mechanism makes it possible to efficiently analyze the titling errors.

5. Applications of CovN/CovD

We apply the proposed metrics to the evaluation of a topic segmentation task on TV broadcast news (TVBN).

5.1. Corpus

We use two corpora, in our experiments. The first one, called *MCS7-14*, contains 86 TVBN shows recorded in the period from the 10^{th} to the 16^{th} of February 2014. Overall, it contains 997 topic segments, which is equivalent to 895 boundaries. The second one, called *MCS5-15*, includes 26 TVBN provided on 26^{th} and 27^{th} January 2015. It contains 297 topic segments, which is equivalent to 271 boundaries.

5.2. Segmentation algorithm

Our topic segmentation baseline system is based on the analysis of lexical distribution, it derived from the *TextTiling* algorithm [10]. A similarity measure is computed between each pair of adjacent blocks[3]. In automatic transcription, sentence boundary detection is not a trivial task, there are neither punctuations nor capital letters. Rather than sentences, units are *breath groups (BG)* which are sequences of words between two pauses in a speech turn. Similarity is computed using a sliding window of size $2K$ between adjacent blocks of *K BGs* along the show. A high similarity value indicates that the two blocks belong to the same topic. Otherwise, those two blocks belong to two different topics. In [11], we propose two approaches to give a weight to each word of the show according to its degree of importance. We consider this algorithm as our *baseline*. The latter is improved by integrating the speaker distribution with lexical distribution in the cohesion computation with an unsupervised approach [12]. Indeed, TVBN shows usually contain an anchor, reporters and guests. If the anchor is generally present along the show, interviewed guests are likely to speak only in a single subject. Combining speaker identification and spoken name detection can further reinforce the cohesion of topically coherent segments. The concept of cohesion applied to terms distribution

[2]Topic titling is a complementary task of the topic segmentation. It consists in giving a title, to each topic segment extracted from *TVBN* shows.

[3]In the original algorithm, a block is constituted of k sentences.

Table 1: *Performances of topic system in terms of F-measure and CovN/CovD*

	Boundaries (nb.)			Segments (nb.)			Segments (dur.)		
	R	*P*	*Fm*	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	*CovN*	$\mathcal{R}_\mathcal{D}$	$\mathcal{P}_\mathcal{D}$	*CovD*
Corpora MCS7-14									
Baseline	70.7	68.1	69.4	58.5	55.2	56.8	68.4	68.4	68.4
Speech cohesion	73.3	73.2	73.2	64.0	63.0	63.5	74.0	74.0	74.0
Diachronic cohesion	73.6	79.1	76.3	64.6	68.1	66.3	75.1	75.1	75.1
Corpora MCS5-15									
Baseline	65.3	60.8	63.0	49.9	45.3	47.4	61.0	60.9	60.9
Speech cohesion	68.6	72.0	70.1	58.9	59.9	59.4	69.4	68.5	69.0
Diachronic cohesion	69.0	72.9	70.9	60.0	66.5	64.7	70.6	70.8	70.7

can be extended to speakers distribution and generalized to the new notion of *speech cohesion*.
Our latest improvement [13] consists of using semantic relations between words in the speech cohesion algorithm. Semantic relations are extracted from *Google News* on the same day as the TVBN show (diachronic corpus). The distance between words is computed by using *word2vec* [14] toolkit and *NWD* [15] distance. We called this system by *Diachronic cohesion*.

5.3. Results and discussion

Table 1 illustrates the performance of our system in terms of *F-measure* (with a tolerance margin of $10s$) and *CovN/CovD* ($\gamma = 85\%$). We can observe that *F-mesure* (Fm) scores are greater than $CovN$ ones. This comes down to the measure *CovN* which considers that a segment is correct if the boundaries of the beginning and the end of the hypothesis are close to those of the reference.
The system performance is improved when taking account of the conjointly distribution of terms and speakers during the computing of cohesion. On the *MCS7-14* corpus, the *baseline* system returns 1056 segments including 583 correct segments. However, the system based on speech cohesion returns 1012 segments, including 638 correct segments. This allows *CovN* and *CovD* to go from 56.8 to 63.5 and from 68.4 to 74.0 respectively. In terms of the number of correct boundaries returned, Fm goes from 69.4 to 73.2. After analysis of the results achieved, we have seen that adding the labels of the speakers in similarity computation allows to refine the segmentation even if there is little repetition of terms within the topic segment. We note that many false segments are placed in the middle of reports, or even many segments before the end of interview are deleted.
The integration of semantic relations reinforces much more the cohesion computation by returning less segments. In total 946 segments are returned of which 644 are correct. Moreover, $\mathcal{P}_D$ went up from 63.0 to 68.1.
Similar tendencies are observed on the corpus *MCS5-15*, *i.e* speech cohesion improves the system performances by detecting new segments and deleting the false alarms. However, semantic relations focus much more on the detection of false segments.
The difference between the *CovN* and *CovD* scores can be explained by the fact that the system is performing better on the long segments. To verify this, we evaluated our system according to the size of the segments (results are indicated in table 2). From an applicative point of view, we assume that long segments are more important to retrieve than short ones, as they convey more information on their topic and are more likely to be re-used after their first live broadcast. Thus, we make distinct evaluations, according to the duration of the reference segments. When analyzing the distribution of the duration of the segments, it appears that they can be easily divided into two sets, where the threshold between short and long segments is set to $30s$.

Table 2: *The performance of the topic system depends of the segment size.*

Corpora MCS7-14				*Corpora MCS5-15*			
Long seg.		Short seg.		Longs seg.		Short seg.	
$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$	$\mathcal{R}_\mathcal{N}$	$\mathcal{P}_\mathcal{N}$
76.1	77.1	27.5	33.3	72.7	71.7	31.4	43.1

The corpus *MCS7-14* contains 761 and 236 long and short segments respectively. While the *MCS5-15* corpus contains 227 and 70 long and short segments respectively.
The system has better performance on long segments and less efficiency when it comes to short segments. Indeed, with the long segments $\mathcal{R}_D = 76.1$ and $\mathcal{P}_D = 77.1$, while for short segments $\mathcal{R}_D = 27.5$ and $\mathcal{P}_D = 33.3$. Similar tendencies are observed on the corpus *MCS5-15*. This is due to the fact that the long segments contain several indications to extract effectively the beginning and the end of the segment, such as speakers distribution, words repetition and semantic relations between words. The short consecutive segments (*i.e* brief informations) must be identified at the first place (*e.g* detecting the jingles), then apply a specific treatment.

6. Conclusion

In this work, we have proposed evaluation metrics of automatic segmentation, which focus on the segment detection and not on the boundary detection. These measures are more adapted for many applications and their evaluation such as topic titling [8], summarization, information retrieval and speaker diarization which are more related to the segments than to the boundaries. Finally, the proposed measures make it possible to analyze the topic segmentation system according to the properties of the segments.

7. References

[1] Beeferman, D., Berger, A., and Lafferty. Statistical models for text segmentation. Machine learning, 34(1-3):177-210, 1999.

[2] Pevzner, L. and Hearst, A Critique and Improvement of an Evaluation Metric for Text Segmentation, Computational Linguistics, 19–36, 2002.

[3] Lamprier, S., Amghar, T., Levrat, B. and Saubion, F. On evaluation methodologies for text segmentation algorithms. In 19th IEEE International Conference on Tools with Artificial Intelligence Greece, 2007

[4] Sitbon, L., Bellot P. Tools and methods for objective or contextual evaluation of topic segmentation. In Proceedings of Language Resources and Evaluation, 2006.

[5] Georgescul, M., Clark, A. and Armstrong, S. An analysis of quantitative aspects in the evaluation of thematic segmentation algorithms. In Proceedings of the 7th Workshop on Discourse and Dialogue, SigDIA, 2006.

[6] Scaiano, M. and Inkpen, D. Getting more from segmentation evaluation. In Proceedings Conference of North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pages, Montreal, Canada, 2012.

[7] Fournier, C., and Inkpen, D. Segmentation Similarity and Agreement. In Proceedings of Human Language Technologies: The Annual Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, USA, 2012.

[8] Fournier, C. Evaluating text segmentation using boundary edit distance. In Proceedings of the 51st Annual Meeting of the Association for Computational, Linguistics, 2013.

[9] Bouchekif, A., Damnati, G., Charlet, D., Camelin, N., and Estève, Y. Title assignment for automatic topic segments in TV broadcast news, IEEE International Conference on Acoustics, Speech and Signal Processing, Shanghai, 2016.

[10] HEARST M. A. (1997). Texttiling : Segmenting text into multi-paragraph subtopic passages. Comput. Linguist., 23(1), 3364.

[11] Bouchekif, A., Damnati, G., and D. Charlet, Intra-content term weighting for Topic Segmentation, in In 39th IEEE International Conference on Acoustics, Speech and Signal Processing, Italy, 2014.

[12] Bouchekif, A., Damnati, G., and D. Charlet, Speech cohesion for topic segmentation of spoken contents, in 15th Annual Conference of the International Speech Communication Association, Singapore, 2014.

[13] Bouchekif, A., Damnati, G., Estève, Y., Charlet, D., and Camelin, N. Diachronic Semantic Cohesion for Topic Segmentation of TV Broadcast News. In 16th Annual Conference of the International Speech Communication Association, Dresden, 2015.

[14] Mikolov, T., Chen, K., Corrado, G. and Dean, J., Efficient estimation of word representations in vector space, arXiv preprint arXiv:1301.3781, 2013.

[15] Vitányi, P. M., Balbach, F. J., Cilibrasi, R. L., and Li, M. Normalized information distance. In Information theory and statistical learning. Springer US, 2009.