Linking written News and TV Broadcast News topic segments with semantic textual similarity

Delphine Charlet, Géraldine Damnati
Orange Labs, Lannion, France
firstname.lastname@orange.com

Abstract

This article explores the task of linking written and audiovisual News, based on the use of semantic textual similarity metrics. It presents a comprehensive study of different linking approaches with various configurations of inter-media or intra-media association. The influence of document length and request length is also explored. It is shown that textual similarity metrics that have proved to perform very well in the context of community question answering can provide efficient News linking metrics, whatever the media association configuration.

Keywords: textual semantic similarity, multimedia linking

1. Introduction

Linking multimedia information has become an important subject from several perspectives, ranging from helping professional journalists in order to perform news analytics to helping end-users to compare various sources of information. Linking can be seen as an input for the design of news exploration tools (e.g. (Bois et al., 2017b) or as an input for the design of efficient search engines in order to be able to retrieve related information. We propose to address the latter use-case and we compare several similarity metrics in various information retrieval configurations.

Multi-modal linking is a domain where semantic similarities are searched among multi-modal documents and/or across modalities. Video hyperlinking is a task in multimedia evaluation campains such as MediaEval (Eskevich et al., 2014) or TRECVid (see *e.g.* (Bois et al., 2017c)). The objective here is to be able to link an anchor (a piece of a BBC program which has been selected by experts as a segment of interest) to other segments defined as targets, that can be extracted from 2,700 hours of programs. This task, similarly to textual semantic similarity tasks, refers to homogeneous data: the objective is to link a fragment to another fragment from the same source. Some other works attempt to link heterogeneous sources but from an alignment perspective (*e.g.* books and movies (Zhu et al., 2015) or video lectures and scientific papers (Mougard et al., 2015)).

In the News domain there has been several studies about linking press articles with other information sources. (Aker et al., 2015) explore linking press articles and comments on the AT corpus (Das et al., 2014) which has been built from article of The Guardian. Linking press articles and Tweets have also been studied (Guo et al., 2013). (Bois et al., 2017a) attempt to build graph representations for News browsing. The authors have collected over a 3 week period a corpus of documents in French including press articles, videos and radio podcasts. Recently, the FrNewsLink corpus (Camelin et al., 2018) has been released allowing several multi-modal linking tasks to be addressed, with heterogeneous data from various sources and of various length. We propose to study the impact of several semantic similarity metrics on the task of retrieving related pieces of information on the basis of a request provided by a source piece of information.

Semantic similarity between texts have been the subject of several recent research challenges, e.g. (Cer et al., 2017) for computing similarity between sentences, or (Nakov et al., 2017) for the ranking of similar questions from a community forum, for a given question as a request. In these challenges, many sophisticated systems have been developed, mainly in a supervised way, when training data was available to learn how similar the paired texts were. Nevertheless, most of the proposed solutions were a supervised combination of unsupervised similarity metrics. Among these unsupervised similarity metrics, one appeared to perform noticeably well, in the case of Community Question Answering: the `soft-cosine` measure, which take advantage of word-embeddings based word relations, in a framework that generalizes the cosine similarity between bag-of-words (Charlet and Damnati, 2017).

In this work, we propose to explore the potential of this metric for linking heterogeneous data, namely written News and topic fragments of audiovisual News, with experiments on the FrNewsLink corpus. Section 2. presents this corpus and the various linking tasks. Section 3. presents the similarity metrics which are evaluated in section 4..

2. Corpus and linking tasks

2.1. Multimedia Corpus

We use *FrNewsLink*, a corpus which is publicly available (Camelin et al., 2018). This corpus contains automatic transcriptions of TV Broadcast News (TVBN) shows, as well as texts extracted from on-line press articles of the same period. The *FrNewsLink* corpus is based on 112 (TVBN) shows from 8 different French channels recorded during two periods in 2014 and 2015. Manual annotation for topic segmentation is provided for TVBN, thus the corpus contains mono-thematic segments of automatic transcriptions of News. For this work, we use the set of TVBN shows that have been collected during the 7^{th} week of 2014, containing 86 news shows from 8 different channels, and yielding an amount of 992 mono-thematic segments.

A set of 24,7k press articles published at the same period

has been gathered. Additionally, manual annotation is provided, that links press articles and TVBN segments. A press article is linked to a TVBN segment if they are both from the same day and if the title of the press article can be considered as an acceptable title for the TVBN segment.

2.2. Inter-media linking

Thanks to this multi-media annotated corpus, we can evaluate different inter-media linking tasks. One is to consider a speech segment as a "request" with the purpose of retrieving all the press articles of the same day that are linked to the segment. Conversely, we can consider a press article as a "request" and the task is to retrieve all the speech segments of the same day linked to the press article. Table 1 gives figures describing the corpus W07_14 (corresponding to 7^{th} week of 2014) in the inter-modal perspective.

# TVBN segments	992
# TVBN segments with at least one linked press article	707
average number of linked press article per segment with at least one linked article	11.1
# press article	5024
# press article with at least one linked TVBN news	1784
average number of linked TVBN segments per press article with at least one segment	4.4
# of inter-media linked pairs (TVBN segment with linked press article of the same day)	7830
# of potential pairs (TVBN segments $\times$ press article of the same day)	734 113
percentage of linked pairs among potential pairs	1.1%

Table 1: W07_14 statistics for inter-media linking

2.3. Intra-media linking

From the above mentioned manually annotated corpus, we can also build 2 intra-media linking tasks, through indirect supervision.

2.3.1. Linking TV Broadcast News segments

We consider that 2 TVBN segments are linked if there exists at least one common press article linked to both segments. If 2 TVBN segments are linked to press articles but without any article in common, we consider that these 2 segments are not linked. We cannot conclude about the existence or the absence of a link between 2 TVBN segments which are not linked to any press article (they could be linked, based on a topic which is not present in the press article corpus). Thus, in order to explore linking between TVBN segments, we restrict the corpus to the set of TVBN segments having at least one linked press article. Table 2 presents the statistics related to this task.

It is worth noticing that among the 707 TVBN segments which have at least one linked press article, 85% of them (604) also have a link with another TVBN segment. Only 15% of the TVBN segments linked with press articles has no other linked segments. It means that a topic from a TVBN show which is also present in written press, is very likely to be addressed in other TVBN shows during the day.

# TVBN segments (*with at least one linked press article*)	707
# TVBN segments linked with at least one TVBN segment	604
average number of linked segments per segment with at least one linked segment	11.3
# of intra-media TVBN segments linked pairs of the same day	6844
# of potential pairs of TVBN segments of the same day	76444
percentage of linked pair among potential pairs	9.0%

Table 2: W07_14 statistics for TVBN segments linking

2.3.2. Linking press articles

Conversely, we can apply the same approach to build a corpus of linked press articles. We consider that 2 press articles are linked if there exists at least one common TVBN segment linked to both articles. If 2 press articles are linked to TVBN segments but without any one in common, we consider that these 2 press articles are not linked. We cannot conclude about the existence or the absence of a link between 2 press articles which are not linked to any TVBN segments (they could be linked, based on a topic which is not addressed in TVBN shows). Thus, in order to explore linking between press articles, we restrict the corpus to the set of press articles having at least one linked TVBN segment. Table 3 shows some statistics if we consider the task of news retrieval for a given press article as a request. We can notice that among the 1784 press articles which have at least one linked TVBN segment, 97% of them (1734) have also a link with another press article. It means that once a topic from on line press is also present on TV, it is highly probable that other press articles deal with the same topic. It emphasizes the phenomenon noticed in previous section about TVBN. Thus, the existence, for a specific topic, of a cross-media link between TV and on line press implies that this topic has a high probability of being treated multiple times within each media.

# press article (*with at least one linked TVBN news*)	1784
# press articles linked with at least one press article	1734
average number of linked articles for an article with at least one linked article	20.8
# of linked pairs of articles of the same day	36126
# of potential pairs of articles of the same day	482132
percentage of linked pairs among potential pairs	7.5%

Table 3: W07_14 statistics for press articles linking

3. Similarity Metrics

3.1. Preprocessing and baseline

The texts are lemmatized and only the lemmas of adjectives, verbs and nouns are selected. Okapi TF-IDF$_{BM25}$ weights are estimated from the aggregated news corpus of the whole week. Hence, text representation consists of a weighted bag of lemmas. As a baseline similarity metrics, a `cosine` similarity is computed between vectors X and Y respectively representing texts T_X and T_Y:

$$cos(X,Y) = \frac{X^t.Y}{\sqrt{X^t.X}\sqrt{Y^t.Y}} \text{with } X^t.Y = \sum_{i=1}^{n} x_i y_i \quad (1)$$

3.2. `soft-cosine` similarity

When there are no words in common between texts T_X and T_Y (*i.e.* no index i for which both x_i and y_i are not equal to zero), `cosine` similarity is null. However, even with no words in common, texts can be semantically related when the words are themselves semantically related. This is why some authors (Sidorov et al., 2014) (Charlet and Damnati, 2017) have proposed to take into account word-level relations by introducing in the `cosine` similarity formula a relation matrix M, as suggested in equation 2.

$$cos_M(X,Y) = \frac{X^t.M.Y}{\sqrt{X^t.M.X}\sqrt{Y^t.M.Y}} \quad (2)$$

$$X^t.M.Y = \sum_{i=1}^{n}\sum_{j=1}^{n} x_i m_{i,j} y_j \quad (3)$$

where M is a matrix whose element $m_{i,j}$ expresses some relation between word i and word j. With such a metric, the similarity between two texts is non zero as soon as the texts share related words, even if they have no word in common. Introducing the relation matrix in the denominator normalization factors ensures that the reflexive similarity is 1.
Here, M reflects the similarity between word embeddings. This metric proved in SemEval2017 to be very efficient to measure similarity between questions in social data in order to address the Community Question Answering task(Charlet and Damnati, 2017). As proposed in the paper, the matrix element $m_{i,j}$ is computed as:

$$m_{i,j} = max(0, cos(v_i, v_j))^2 \quad (4)$$

where v_i and v_j are the embeddings for words i and j. They are estimated with word2vec tool (Mikolov et al., 2013) on the whole corpus of press articles of the given week.

3.3. Text Embeddings

A very simple yet efficient representation for texts consists in simply averaging the embeddings of the words of the text. It was used by many participants of the last SemEval challenge on Community Questions Answering (Nakov et al., 2017). Weighting the contribution of each word in the average embeddings appears to give significant improvement and to be competitive compared with other methods of text embeddings (Arora et al., 2017). Thus, if x_i is the weight of word i and $v_{i,k}$ is the k^{th} component of word i in the embeddings space, the k^{th} component of vector $\tilde{X}$, which represents text T_X is:

$$\tilde{X}_k = \frac{1}{\sum_i x_i} \sum_i x_i v_{i,k}$$

The similarity measure `wavg-w2v`(X,Y) between T_X and T_Y is then computed as the cosine between $\tilde{X}$ et $\tilde{Y}$.

4. Experiments

4.1. Protocol and evaluation metrics

We adopt a general protocol, whatever the type of texts to be linked (TVBN segments or press article). The task of retrieving, for a given request, the linked texts, is evaluated with Mean Average Precision (MAP). For a given request, similarities between the request and all the potential texts of the same day are computed, and the texts are ranked according to decreasing similarity. MAP@10 is used to evaluate the pertinence of the ranking of the 10 most similar texts. MAP measures how well the texts that should be linked (based on the ground-through annotations), are ranked before the non-linked texts, when the ranking is based on textual similarity metric.

As a complement to Information Retrieval evaluation, we can also consider the task of detecting linked pairs, among all potential pairs. Similarities are computed between all potential pairs of texts, and those whose similarity is above a certain threshold are considered as linked. For this set of detected pairs, precision and recall rates can be computed, as well as their harmonic mean, the F-measure. MAP and F-measure reflect different points of view: MAP translates the ability of the similarity metrics to rank the linked texts before the non-linked texts, for a given text request, without any notion of decision threshold. The F-measure additionally measures the ability to set a threshold on the similarity value in order to decide if a pair is linked or not, this threshold being common to all requests. Beyond the ability to rank, a good F-measure reflects the fact that similarity metrics between different text pairs are comparable. In the tables presented in the next section, the threshold is set *a posteriori*, so as to get the maximal F-measure.

Press articles are composed of a title (the first line of the extracted text) and a body (the rest of the text). Contrastive experiments are systematically done, considering either the title or the full article, to compute the similarity metrics. In fact, text length of the elements to be linked is expected to have a major influence on performances. Text length is given in terms of different selected lemmas, which is the size of the bag-of-words vector we keep. The average length of TVBN segments is 42.1 words. For press article, the average length of titles is 6.6 words, whereas the average length of the full articles is 120.8 words.

4.2. Results

The first set of experiments reflect the condition where the request is a TVBN segment. Table 2 presents the results for TVBN segments linking. The task is to retrieve the TVBN segments that address the same topic as the request TVBN segment. `soft-cosine` and `wavg-w2v`, which are the

request: TVBN segment		MAP@10	Fmax
target: TVBN segments	cosine	0.896	61.5
	soft-cosine	**0.932**	66.8
	wavg-w2v	**0.930**	**68.0**

Table 4: TVBN segments intra-linking

request: TVBN segment		MAP@10	Fmax
target: press article title	cosine	0.680	53.7
	soft-cosine	**0.750**	**66.1**
	wavg-w2v	0.743	65.1
target: full press article	cosine	**0.834**	73.5
	soft-cosine	0.820	**74.5**
	wavg-w2v	0.807	72.1

Table 5: linking press articles to TVBN segment

metrics which exploit word embeddings, perform equivalently (with an advantage to wavg-w2v for Fmax) and significantly better than the cosine metric. One can notice that, if the MAP@10 gives pretty good performance, the F-measure, which also involves a common decision threshold on the metrics, is not so good.

Then, table 5 presents the results for linking TVBN segments to press articles, with 2 variants: first, the press article is only represented by its title, second, the entire press article is considered. When it comes to link TVBN segments towards very short texts (titles), performances are worse than when the full article is considered. Interestingly, while the metrics which use word-embeddings perform much better than the bag-of-words cosine for linking towards titles, it is not the case for linking towards longer texts. Indeed, the bag-of-words cosine obtains the best MAP in this case. It is consistent with the fact that the advantage of word-embeddings based metrics is to measure a similarity even between texts without any word in common. The shorter the texts, the more likely it is that they do not share any word. When documents are long, it is very probable that they have common words if they are related. In this case, introducing semantic relations between words in the metrics can yield some noise by inducing too many relations and the metrics loose in their ability to rank target articles. However, when considering F-max, soft-cosine performs the best whatever the length of the target. This suggests that this metric remains better in its ability to set a threshold for detecting similar pairs, as the similarity values are more consistent.

Table 6 presents results obtained for the symmetric task: the request is now the press article, which has to be linked to TVBN segments. The variants considering only the title or the entire press article as request are evaluated. Here again, we can observe that performances obtained when using the full article are better than the ones obtained with the title only. It is also the cosine metric which is the most sensitive to increasing request length: MAP raises from 0.761 with the title to 0.917 with the full article. wavg-w2v does not benefit a lot from the increasing of text available in the

request: press article title		MAP@10	Fmax
target: TVBN segments	cosine	0.761	53.7
	soft-cosine	0.**889**	**66.1**
	wavg-w2v	**0.887**	65.1
request: full press article		MAP@10	Fmax
target: TVBN segments	cosine	0.917	73.5
	soft-cosine	**0.923**	**74.5**
	wavg-w2v	0.896	72.1

Table 6: linking TVBN segments to press articles

request: press article title		MAP@10	Fmax
target: press article title	cosine	0.807	56.2
	soft-cosine	0.865	58.6
	wavg-w2v	**0.910**	**75.4**
target: full press article	cosine	0.907	67.7
	soft-cosine	**0.941**	**80.3**
	wavg-w2v	0.933	0.79.3
request: full press article		MAP@10	Fmax
target: press article title	cosine	0.918	67.7
	soft-cosine	**0.933**	**80.3**
	wavg-w2v	0.911	79.3
target: full press article	cosine	**0.940**	78.9
	soft-cosine	**0.939**	**83.7**
	wavg-w2v	0.927	82.3

Table 7: Press articles intra-linking

request: its MAP starts at 0.887 with the title, and ends at 0.896 with the full article. soft-cosine performs best, whatever the length of the request.

Finally, table 7 presents the results for intra-media press article linking, with all the possible variants, whether the request or the target are built with the title only or the full article. When it comes to link short texts together (title/title), wavg-w2v performs significantly better than the other metrics. soft-cosine, which also used word-embeddings, performs better than cosine, but not as good as wavg-w2v. When it comes to link long texts together (full/full), cosine and soft-cosine obtain the best MAP. It is worth noticing that for an equivalent MAP, cosine get a far worse F-max than the alternative metrics. For instance, in the case of (full/title), cosine and wavg-w2v get a MAP around 0.91 but a F-max respectively of 67.7 and 79.3. Likewise, in the case of (full/full), cosine and soft-cosine get a MAP around 0.94 but a F-max respectively of 78.9 and 83.7. It means that cosine provides scores which are good for ranking, but not as good as the other metrics for a global decision threshold. When it comes to link content of very different lengths (full/title or title/full), the soft-cosine performs the best.

5. Conclusion

We have shown in this article that similarity metrics that have proved to perform well for social media linking can be also very efficient for multi-media News linking. We have proposed a comprehensive study which presents the pros and cons of three different metrics in various inter-

media and intra-media linking configuration. We can draw some specific conclusions about the metrics. The baseline bag-of-words `cosine` is the most sensitive to the length of texts to be linked. It performs the worst for very short texts, and the best, or close to the best, for long texts. `wavg-w2v` is the best metric by far when it comes to linking very short texts together (title/title), but as soon as there is a mismatch in the size of texts to be linked, `soft-cosine` is better. In our experiments, the obtained MAP are pretty high, but there is a lot of room for improvement for F-max. It means that further work is necessary to make the metrics more robust to global decision threshold.

6. Bibliographical References

Aker, A., Kurtic, E., Hepple, M., Gaizauskas, R., and Di Fabbrizio, G. (2015). Comment-to-article linking in the online news domain. In *Proceedings of the SIGDIAL 2015 Conference*, pages 245–249. ACL.

Arora, S., Liang, Y., and Ma, T. (2017). A simple but tough-to-beat baseline for sentence embeddings. In *Proceedings of ICLR 2017*, Toulon, France, April.

Bois, R., Gravier, G., Jamet, É., Morin, E., Robert, M., and Sébillot, P. (2017a). Linking multimedia content for efficient news browsing. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval, ICMR 2017, Bucharest, Romania, June 6-9, 2017*, pages 301–307.

Bois, R., Gravier, G., Jamet, E., Morin, E., Sébillot, P., and Robert, M. (2017b). Language-based construction of explorable news graphs for journalists. In *Proceedings of the 2017 EMNLP Workshop: Natural Language Processing meets Journalism*, pages 31–36, Copenhagen, Denmark, September. Association for Computational Linguistics.

Bois, R., Vukotić, V., Simon, A.-R., Sicre, R., Raymond, C., Sébillot, P., and Gravier, G., (2017c). *Exploiting Multimodality in Video Hyperlinking to Improve Target Diversity*, pages 185–197. Springer International Publishing, Cham.

Camelin, N., Damnati, G., Bouchekif, A., Landeau, A., Charlet, D., and Estève, Y. (2018). Frnewslink : a corpus linking tv broadcast news segments and press articles. In *Proceedings of LREC 2018*, Miyazaki, Japan, May.

Cer, D. M., Diab, M. T., Agirre, E., Lopez-Gazpio, I., and Specia, L. (2017). Semeval-2017 task 1: Semantic textual similarity multilingual and crosslingual focused evaluation. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 1–14.

Charlet, D. and Damnati, G. (2017). Simbow at semeval-2017 task 3: Soft-cosine semantic similarity between questions for community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation, SemEval@ACL 2017, Vancouver, Canada, August 3-4, 2017*, pages 315–319.

Das, M. K., Bansal, T., and Bhattacharyya, C. (2014). Going beyond corr-lda for detecting specific comments on news & blogs. In *Proceedings of the 7th ACM international conference on Web search and data mining*, pages 483–492. ACM.

Eskevich, M., Aly, R., Racca, D., Ordelman, R., Chen, S., and Jones, G. J. (2014). The search and hyperlinking task at mediaeval 2014.

Guo, W., Li, H., Ji, H., and Diab, M. T. (2013). Linking tweets to news: A framework to enrich short text data in social media. In *ACL (1)*, pages 239–249.

Mikolov, T., Sutskever, I., Chen, K., Corrado, G. S., and Dean, J. (2013). Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*.

Mougard, H., Riou, M., de la Higuera, C., Quiniou, S., and Aubert, O. (2015). The paper or the video: Why choose? In *Proceedings of the 24th International Conference on World Wide Web*, pages 1019–1022. ACM.

Nakov, P., Hoogeveen, D., Màrquez, L., Moschitti, A., Mubarak, H., Baldwin, T., and Verspoor, K. (2017). SemEval-2017 task 3: Community question answering. In *Proceedings of the 11th International Workshop on Semantic Evaluation*, SemEval '17, Vancouver, Canada, August. Association for Computational Linguistics.

Sidorov, G., Gelbukh, A. F., Gómez-Adorno, H., and Pinto, D. (2014). Soft similarity and soft cosine measure: Similarity of features in vector space model. *Computación y Sistemas*, 18(3).

Zhu, Y., Kiros, R., Zemel, R., Salakhutdinov, R., Urtasun, R., Torralba, A., and Fidler, S. (2015). Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE international conference on computer vision*, pages 19–27.