Automatic Sentence Ordering Assessment Based on Similarity

Liana Ermakova
IRIT
Toulouse, France
ermakova@irit.fr

ABSTRACT

One of the tasks of text generation is sentence ordering since it is crucial for readability. Nevertheless, there is no common approach for evaluation of sentence ordering. The state-of-the art methods are based on the comparison with a human-provided order. However, in many cases it is impossible or time and resource consuming. Therefore, we propose three completely automatic approaches for sentence order assessment where the similarity between adjacent sentences is used as a measure of text coherence. We showed that the methods based on word and noun similarities have very high agreement with the human-provided judgment. We also propose an automatic evaluation framework for analysis of the metrics of sentence order that requires only a text collection.

CCS Concepts

•Information systems → Information retrieval; Evaluation of retrieval results;

Keywords

Information retrieval, evaluation, text generation, sentence ordering, similarity

1. INTRODUCTION

One of the tasks of text generation is sentence ordering. Barzilay *et al.* showed that sentence order influences a lot text perception [2]. To evaluate the sentence order produced by text generation algorithms an assessment metrics is needed. The majority of techniques involve human intervention. The common practice is to evaluate sentence order by rank correlation coefficient between a gold standard and a candidate [8]. However, manual judgment is expensive and time consuming and, thus, not applicable in real time or on large collections. In some cases (e.g. comparison with a pool of reference sentences) the use of rank correlation coefficients is impossible since there is no gold standard order. Moreover, since correct sentence order may be not unique, providing all possible correct orders may be expensive. That is why the main objective of this research is to provide completely automatic methods for evaluation of sentence ordering.

Our hypothesis is that a good sentence order implies the similarity between adjacent sentences since word repetition is one of the formal sign of text coherence [2]. Therefore, we propose three approaches to assess the global coherence of a text on the basis of its graph model, where the vertices correspond to sentences and the edges represent the similarity measure between them. Our methods are based on the similarities of terms, nouns and named entities. In this case, sentence order is viewed as the path passing through each vertex exactly once.

The second contribution of this paper is an automatic evaluation framework for analysis of the metrics of sentence order that requires only a text collection.

The rest of the paper is organized as follows. Section 2 provides the state of the art in sentence ordering and its evaluation. In Section 3 we describe three automatic methods we promote to sentence ordering assessment. In Section 4 we propose an automatic framework to evaluate assessment metrics of sentence ordering. Section 5 presents the obtained results and discusses them. Section 6 concludes the paper.

2. STATE OF THE ART

2.1 Sentence Ordering

In single document summarization the sentence order may be the same as the initial relative order in the original text. However, this technique is not applicable to multi-document summarization. The retrieved sentences should be organized into a coherent text. If an extraction system deals with entire passages, locally they may have higher readability than generated phrases since they are written by humans. Nevertheless, it is important to keep in mind the global readability of extracted passages. The only way to improve the readability of a text produced by an extraction system is to reorder the extracted passages.

According to Centering Theory there are four types of links between sentences: Continuation > Retaining > Smooth Shift > Rough Shift [6].

The idea of using the initial sentence order was adopted by Majority Ordering algorithm for multi-document summarization. In this approach subjects (sentences expressing the same meaning) T_i are organized into a directed graph were edges present the number of documents where T_i is

EVIA '16, Tokyo, Japan
© 2016 ACM. ISBN 123-4567-24-567/08/06...$15.00
DOI: 10.475/123_4

followed by T_j and the best order corresponds to the Hamiltonian path of maximal length [2]. Another approach is to assign time stamp to each event and to order them chronologically. The use of chronological ordering is restricted to the news articles on the same topic [2]. Diversity topics in the news demand another way to arrange sentences extracted for multi-document summarization. Application of a text corpus provides the ground for improving readability. In this case the optimal order is found by the greedy algorithm maximizing the total probability [8]. In a narrative text verbs and adjectives play an important role in the semantic relations between sentences [1]. Specific ordering is applied to verb tenses [8]. Lapata proposed to consider sentence ordering as a Markov process [8]. In [21] sentence re-ordering is based on document time-stamps and sentence position within a document. No further details are provided. Ying et al. re-ordered sentences according to its timestamps in the original document [20]. To ensure local coherence of a summary, Hickle et al. used a hierarchical clustering algorithm to re-order sentences that contain similar information [7]. Mihalcea and Tarau suggested a directed backward graph where the edges are oriented from a sentence to previous sentences in the text [10].

2.2 Evaluation

Traditional methods of readability evaluation are based on the familiarity of terms and syntax complexity [4, 12, 19, 22]. Syntactical methods may be combined with statistics (e.g. sentence length, the depth of a parse tree, omission of personal verb, rate of prepositional phrases, noun and verb groups etc.) [3]. Last methods are suitable only for the readability evaluation of a particular sentence and therefore they cannot be used for extracts assessment.Researches also propose to use language models [4, 15]. Usually assessors assign score to the readability of text in some range [2]. Syntactical errors, unresolved anaphora, redundant information and coherence influence readability and therefore the score may depend on the number of these mistakes [14].

Traditionally, evaluation of sentence order is based on the comparison of the ranked lists of sentences in a gold standard and a candidate [8]. Different non-parametric rank correlation coefficients (e.g. Kendall, Spearman or Pearson coefficients) may be used to find the dependence [9]. However as it is shown in [8], Kendall coefficient is the most suitable for sentence ordering assessment:

$$\tau = 2 \times \frac{(number(agreement) - number(disagrement))}{N(N-1)} \tag{1}$$

Since sentence ordering may be “correct”, but not necessary unique, it is advisable to consider an average value of different results [8]. BLEU and edit distance may be applied for relevance judgment as well as for readability evaluation. Like correlation coefficients, these metrics are semi-automatic since they require gold standards.

3. PROPOSITION

We propose three approaches to evaluate the global coherence of text (mainly sentence ordering) on the basis of its graph model , where vertices represent sentences and edges correspond to the similarity measure between them. The hypothesis is that neighboring sentences should be somehow similar to each other. Thus, we introduce here three methods for sentence ordering assessment based on similarity measures between the sentences adjacent in the text. The major advantages of the proposed approaches is that they are completely automatic and they do not require any external source.

3.1 Similarity for Sentence Ordering Assessment

In the first model we propose, the score of a document D is computed as the normalized similarity between all adjacent sentences (S_i, S_{i+1}):

$$score(D) = \frac{\sum_{i=1}^{n-1} sim(S_i, S_{i+1})}{n-1} \tag{2}$$

where n is the total number of sentence in the document D.

The similarity between sentence is estimated by the cosine measure:

$$sim(S,P) = cos(S,P) = \frac{\sum_{i=1}^{n} S_i P_i}{\sqrt{\sum_{i=1}^{n} S_i^2}\sqrt{\sum_{i=1}^{n} P_i^2}} \tag{3}$$

where S and P are the adjacent sentences, S_i and P_i are the $i-$th terms in the vector representation of the sentences S and P.

This method requires only stemming and sentence chunking.

3.2 Nouns

We adopted the idea of H. G. Silber and K. F. Mccoy that nouns provide the most valuable information [16]. Therefore, we map each sentence into to the set of nouns. Then we calculate the cosine similarity measure between the adjacent sentences. The total score of a text is estimated as an average noun set cosine similarity between sentences. The method requires par-of-speech tagging.

3.3 Named Entities

Many researches indicate that named entity recognition may improve information retrieval performance, including tweet study [5, 11, 13]. Thus, we decided to try to apply named entity similarity for the sentence ordering evaluation. The total score of a document is calculated as a normalized named entity similarity. As in two previous approaches, we use the cosine similarity between adjacent sentences.

4. EVALUATION FRAMEWORK

4.1 Source Data Sets

For test collection generation we used three source datasets:

- TREC (Text Retrieval Conference) Robust data set;
- WT10G;
- Wikipedia dump.

TREC **Robust** Track data set is a ”pure” collection since the documents have almost the same format and there is no spam [18]. There are 4 sources of documents: the news articles from

- The Financial Times, 1991-1994 (FT) - 564MB, 210,158 documents;
- Federal Register, 1994 (FR94) - 395MB, 55,630 documents;

- Foreign Broadcast Information Service (FBIS) - 470MB, 130,471 documents;
- The LA Times - 475MB, 131,896 documents.

Documents are tagged by SGML. The collected documents are not normalized and may contain spelling or other errors. We used 193,022 randomly chosen documents.

In contrast, **WT10G** is a snapshot of the web with real documents in HTML format, some of which are spam. As it was showed in [17], WT10G ”looks like” the web. WT10G was used at TREC Web track 2000-2001. It is 10GB subset of the web snapshot of 1997 from Internet Archive. WT10G contains 1,692,096 documents from 11,680 servers (minimum 5 documents per server). As in the Robust collection, documents are not normalized and tagged by SGML parser. We randomly selected 88,879 documents.

The third data set we used was the recent (April 2011) cleaned English **Wikipedia** XML dump, totally 3,217,015 non-empty pages [14]. All notes, history and bibliographic references were removed. Thus, a page was composed of a title (*title*), an abstract (a) and sections (s). A section had a header ((h)). Abstract and sections contained paragraphs (p) and entities (t) referring to other pages. We chose randomly 32,211 articles.

4.2 Test Collection Construction

We assumed that the best sentence order is produced by a human (gold standard) and a good evaluation measure should small degradation of results provoked by small permutation in sentence ordering and greater rate of shuffling should provoke larger effect since in this case the obtained order is remoter from the human-made one.

Therefore, for evaluation of the proposed measures we suggest the following types of datasets:

- Source collection;
- *Rn*-collection;
- *R*-collection.

Rn-collection is generated from the source collection by a random shift of n sentences within each document. We used $R1$ and $R2$ collections.

R-collection is derived from the source collection by shuffling of sentences within each document.

In average, the degrees of the permutation should to be ordered as follows: $Source < R1 < R2 < ... < Ri < ... < R$. Thus, the scores of a good metric in average should be order like $Source > R1 > R2 > ... > Ri > ... > R$.

This approach for evaluation of sentence ordering is completely automatic and it requires only a text corpus.

5. RESULTS

Table 1 provides the results of the evaluation of the proposed measures and the relative differences of the scores of permuted texts with the scores of the original texts. In this table the column *Method* refers to one of the proposed methods. The column *Score* shows the values obtained by the proposed methods. *R1-score* presents the score obtained for the texts where one sentence was randomly shifted. *R2-score* refers to the values corresponding to texts with two shifted sentences. *R-score* demonstrates the results for the texts with shuffled sentences. *Sim* denotes the method based on the similarity based on all words. *NounSim* and *NESim* take into account only nouns and named entities respectively.

Table 1: General results

Method	Collection	Score	R1-score	R2-score	R-score
Sim	Wikipedia	0.1052	0.1008	0.0977	0.0721
			4.18%	7.44%	33.88%
	Robust	0.1032	0.0988	0.0959	0.074
			4.26%	7.39%	30.45%
	WT10G	0.1024	0.0978	0.0943	0.0592
			4.49%	8.28%	45.81%
NounSim	Wikipedia	0.0964	0.0914	0.0882	0.0599
			5.19%	8.97%	41.38%
	Robust	0.1078	0.1028	0.0991	0.0729
			4.64%	8.46%	35.22%
	WT10G	0.0852	0.0813	0.0783	0.0448
			4.58%	8.49%	51.60%
NESim	Wikipedia	0.0178	0.0187	0.0191	0.0205
			-5.06%	-6.95%	-14.14%
	Robust	0.0102	0.01	0.0098	0.0092
			1.96%	4.00%	10.20%
	WT10G	0.0275	0.0266	0.0259	0.0228
			3.27%	6.02%	18.15%

For the Wikipedia articles, TREC and WT10G documents according to the methods *Sim* and *NounSim* $score >> R1-score >> R2-score >> R-score$ and the difference is significant at $p = 0.05$ (Student t-test). This fact proves that the proposed measures agree with the human judgments and that smaller permutations in sentence order provoke smaller changes in the score. Therefore, we can conclude that the measures *Sim* and *NounSim* are quite good for sentence order evaluation.

For TREC and WT10G we can observe the same tendency for the method *NESim*. However, at the Wikipedia collection the method *NESim* showed the inversed trend. Thus, it is possible to draw a conclusion that *NESim* is not a universal metric. Moreover, it is less stable since NE may be quite sparse. Since many sentences do not have NE, the similarity between them is often 0.

The table provides evidence that *NounSim* demonstrates the higher difference between human-made and shuffled texts than *Sim* for all test collections. It allows to state that *NounSim* should be preferred over *Sim* in case when one want to differentiate more the obtained results.

One of the drawbacks of the proposed models is that the assigned the same score to the inversed orders (e.g. order $A - B - C$ obtains the same score as $C - B - A$). This fact is caused by the symmetric similarity measure we use (cosine). Although, it is quite easy to cope this problem by applying a non-symmetric similarity or divergence (e.g.

Kullback-Leibler).

The main disadvantage of the methods described in this paper is that there is a possibility to fit to this metrics for example by modeling sentence ordering as a traveling salesman problem. However, we believe that any completely automatic metric suffers from this shortcoming. Moreover, fitting to our metric is NP-hard problem while our algorithm is linear over the number of sentences in a text.

6. CONCLUSIONS

Nowadays there is no common approach for evaluation of sentence ordering though there are several metrics of quality assessment. Many techniques involve human intervention. Manual evaluation is expensive and subjective and it is not applicable in real time or on a large corpus.

In this paper we proposed an evaluation framework for analysis of the metrics of sentence order. The proposed approach is completely automatic and it requires only a text collection.

We introduced three automatic methods for evaluation of sentence order within a text that are based on the similarity between adjacent sentences. The proposed methods have linear complexity over the number of sentences in a text. We evaluated our methods on three test collections. We showed that the methods based on word and noun similarities have very high agreement with the human-provided judgment. We believe that the main disadvantage of our methods is common for all completely automatic metric.

In future we plan to enrich our model a weighting for the parts of speech. It seems useful to analyze non-symmetric similarity. One of the promising direction of the future work seems to be the integration of co-reference resolution.

7. REFERENCES

[1] N. Asher and A. Lascarides. *Logics of Conversation.* Cambridge University Press, 2003.

[2] R. Barzilay, N. Elhadad, and K. R. McKeown. Inferring strategies for sentence ordering in multidocument news summarization. *Journal of Artificial Intelligence Research*, pages 35–55, 2002. 17.

[3] J. Chae and A. Nenkova. Predicting the fluency of text with shallow structural features: case studies of machine translation and human–written text. *Proceedings of the 12th Conference of the European Chapter of the ACL*, pages 139–147, 2009.

[4] K. Collins-Thompson and J. Callan. A language modeling approach to predicting reading difficulty. *Proceedings of HLT/NAACL*, 4, 2004.

[5] D. M. de Oliveira, A. H. Laender, A. Veloso, and A. S. da Silva. Fs-ner: A lightweight filter-stream approach to named entity recognition on twitter data. In *Proceedings of the 22Nd International Conference onArabic named entity recognition World Wide Web Companion*, WWW '13 Companion, pages 597–604, Republic and Canton of Geneva, Switzerland, 2013. International World Wide Web Conferences Steering Committee.

[6] B. Grosz, A. Joshi, and S. Weinstein. Centering: A framework for modeling the local coherence. *Computational Linguistics*, 2(22):203–225, 1995.

[7] A. Hickl, K. Roberts, and F. Lacatusu. LCCâĂŹs GISTexter at DUC 2007: Machine reading for update summarization. In *Proc. of DUC*, volume 7, 2007.

[8] M. Lapata. Probabilistic text structuring: Experiments with sentence ordering. *Proceedings of ACL*, pages 542–552, 2003.

[9] G. Lebanon and J. Lafferty. Cranking: Combining rankings using conditional probability models on permutations. *Machine Learning: Proceedings of the Nineteenth International Conference*, pages 363–370, 2002.

[10] R. Mihalcea. Graph-based ranking algorithms for sentence extraction, applied to text summarization. In *Proceedings of the ACL 2004 on Interactive poster and demonstration sessions*, page 20. Association for Computational Linguistics, 2004.

[11] N. Mohammed and N. Omar. Arabic named entity recognition using artificial neural network. 8(8):1285–1293, 2012.

[12] A. Mutton, M. Dras, S. Wan, and R. Dale. Gleu: Automatic evaluation of sentence–level fluency. *ACL'07*, pages 344–351, 2007.

[13] D. Nadeau and S. Sekine. A survey of named entity recognition and classification. *Lingvisticae Investigationes*, 30(1):3–26, 2007.

[14] E. SanJuan, V. Moriceau, X. Tannier, P. Bellot, and J. Mothe. Overview of the INEX 2011 question answering track (QA@INEX). In S. Geva, J. Kamps, and R. Schenkel, editors, *Focused Retrieval of Content and Structure*, volume 7424 of *Lecture Notes in Computer Science*, pages 188–206. Springer Berlin Heidelberg, 2012.

[15] L. Si and J. Callan. A statistical model for scientific readability. *Proceedings of the tenth international conference on Information and knowledge management*, pages 574–576, 2001.

[16] H. G. Silber and K. F. Mccoy. Efficiently computed lexical chains as an intermediate representation for automatic text summarization. *Computational Linguistics – Summarization*, 28(4):1–11, 2002.

[17] I. Soboroff. Does wt10g look like the web? In *Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '02, pages 423–424, New York, NY, USA, 2002. ACM.

[18] E. M. Voorhees and D. Harman. *Overview of the Sixth Text REtrieval Conference (TREC–6)*. 2000.

[19] S. Wan, R. Dale, and M. Dras. Searching for grammaticality: Propagating dependencies in the viterbi algorithm. *Proceedings of the Tenth European Workshop on Natural Language Generation*, 2005.

[20] J. C. Ying, S. J. Yen, Y. S. Lee, Y. C. Wu, and J. C. Yang. Language model passage retrieval for question-oriented multi document summarization. In *Proc. of Document Understanding Conference*, 2007.

[21] J. Zhang, H. Xu, X. Wang, H. Shen, and Y. Zeng. ICT CAS at DUC 2007. In *Proceedings of the Document Understanding Conference 2007*, 2007.

[22] S. Zwarts and M. Dras. Choosing the right translation: A syntactically informed classification approach. *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 1153–1160, 2008.