A Metric for Sentence Ordering Assessment Based on Topic-Comment Structure

Liana Ermakova, Josiane Mothe, Anton Firsov

A Metric for Sentence Ordering Assessment Based on Topic-Comment Structure

Liana Ermakova
LISIS, CNRS-ESIEE-INRA-UPEM
Université de Lorraine
5 boulevard Descartes
Champs-sur-Marne, France 77454
liana.ermakova@univ-lorraine.fr

Josiane Mothe
IRIT, URM5505 CNRS
ESPE, Université de Toulouse
118 route de Narbonne
Toulouse, France 31062
josiane.mothe@irit.fr

Anton Firsov
Perm State University
15 Bukireva st.
Perm, Russia 614990
a_firsov@mail.ru

ABSTRACT

Sentence ordering (SO) is a key component of verbal ability. It is also crucial for automatic text generation. While numerous researchers developed various methods to automatically evaluate the informativeness of the produced contents, the evaluation of readability is usually performed manually. In contrast to that, we present a self-sufficient metric for SO assessment based on text topic-comment structure. We show that this metric has high accuracy.

KEYWORDS

Information retrieval, evaluation, text coherence, sentence ordering, topic, comment, information structure, topic-comment structure

1 INTRODUCTION

Sentence order (SO) has a strong influence on text perception and understanding [1]. Let consider the following example:

Example 1.1. The Nibelung is the dwarf Alberich, and the ring in question is the one he fashions from the Rhine Gold. Wagner's opera title *Der Ring des Nibelungen* is most literally rendered in English as *The Ring of the Nibelung*.

The text is hardly comprehensible. When we are reading *the Nibelung* or *the ring in question*, we are asking ourselves *what's it all about? which Nibelung? what question?* even if in the next sentence it becomes clearer. Let us now reverse the two sentences:

Example 1.2. Wagner's opera title *Der Ring des Nibelungen* is most literally rendered in English as *The Ring of the Nibelung*. The Nibelung is the dwarf Alberich, and the ring in question is the one he fashions from the Rhine Gold.

Now, it is clear that *the Nibelung* and *the ring in question* explain the opera title *The Ring of the Nibelung*. These examples illustrate that appropriate SO is crucial for readability. Automatic text generation, particularly multi-document extractive summarization, systems also face SO problem [3, 4]. We distinguish two of sentences and (2) to measure the order quality, i.e. determining how well a given list of sentences is ordered. In this paper, we focus on the latter task. In order to measure the quality of SO, if a gold standard is available, then correlation between the ground truth SO and the system order (e.g. Kendall or Spearman rank correlation coefficients) can be used [19]. The requirement of a gold standard containing the correct SO limits the usage of such methods. Indeed, gold standard is often not available or not obvious to build manually. In contrast, in this paper we propose a self-sufficient metric for text coherence assessment that does not require additional data. We evaluate the quality of the metric using the framework proposed in [12]. To evaluate the text coherence, we use a linguistic approach based on the topic-comment structure of the text and inter-sentence similarity.

*A clause-level **topic** is the phrase in a clause that the rest of the clause is understood to be about, and the **comment** is what is being said about the topic.*

According to [24], the topic does not provide new information but connects the sentence to the context. Thus, the *topic* and the *comment* are opposed in terms of the given/new information. The contraposition of the given/new information is called **information structure** or **topic-comment structure**. Going back to Example 1.1, *the Nibelung* and *the ring in question* from the first sentence are expected to be already known by the reader, i.e. they represent topics. However, only the next sentence provides the necessary information. In contrast, in Example 1.2 the first mention of *the ring* and *the Nibelung* was given at the end of the first sentence (*The Ring of the Nibelung*) and then is detailed in the second sentence. In the first sentence, *Wagner's opera title Der Ring des Nibelungen* incarnates the **topic** and *is most literally rendered in English as The Ring of the Nibelung* corresponds to the **comment**. In the second sentence, *The Nibelung* and *the ring in question* refers to **topic**, while the **comment** parts are presented by *is the dwarf Alberich* and *is the one he fashions from the Rhine Gold.*

Although, in literature topic-comment structure has been exploited for document re-ranking [13], classification [5], and text summarization [11], to our knowledge, it has never been applied for SO. The contribution of this paper is a completely automatic approach for SO evaluation based on topic-comment structure of a text that requires only shallow parsing and has linear complexity. Our metric considers the pairwise term similarities of the topics and the comments of the adjacent sentences in a text since word repetition is one of the formal signs of text coherence [1].

DOI: http://dx.doi.org/10.1145/3077136.3080720

2 STATE OF THE ART

Current methods to evaluate readability are based on the familiarity of terms and syntax complexity [8]. Word complexity may be estimated by humans [7, 14, 29] or according to its length [30]. Researches also propose to use language models [8, 27]. Usually assessors assign a score to the readability of a text in some range [1]. Syntactical errors, unresolved anaphora, redundant information and coherence influence readability and therefore the score may depend on the number of these mistakes [26]. BLEU and edit distance may be applied for relevance judgment as well as for readability evaluation. These metrics are semi-automatic because they require a gold standard. Another set of methods is based on syntax analysis which may be combined with statistics (e.g. sentence length, depth of a parse tree, omission of personal verbs, rate of prepositional phrases, noun and verb groups) [6, 25, 32, 33], but they remain suitable only for the readability evaluation of a particular sentence and, therefore, cannot be used for assessing extracts. Lapata applies the greedy algorithm maximizing the total probability on a text corpus as well as using a specific ordering to verb tenses [18]. Louis and Nenkova use a hidden Markov model in which the coherence between adjacent sentences is viewed as transition rules between different topics [23]. Barzilay and Lapata introduce an entity grid model where sentences are mapped into discourse entities with their grammatical roles [2]. Entity features are used to compute the probability of transitions between adjacent sentences. Then machine learning classifiers are applied. Elsner and Charniak add co-reference features [10]. Lin et al. ameliorate the model by discourse relations [21]. The entity grid model and its extensions require syntactical parsing. The disadvantages of these models are data sparsity, domain dependence and computational complexity. The closest work to ours is [12] that proposes an automatic approach for SO assessment where the similarity between adjacent sentences is used as a measure of text coherence. However, it assigns equal scores to initial and inverse SO due to the symmetric similarity measure. In contrast, our topic-comment based method assigns higher score to the text in Example 1.2 than 1.1.

3 TOPIC-COMMENT STRUCTURE FOR SO

Although it is not the core element of our method, in order to better understand the topic-comment structure of texts of different genres, we manually examined 10 documents randomly chosen from three datasets (30 texts in total): (1) Wikipedia; (2) TREC Robust[1]; (3) TREC WT10G (for collection details see Section 4). We looked at topic-topic (TT), comment-topic (CT), topic-comment (TC) and comment-comment (CC) inter-sentence relations in the texts, i.e. how frequently a topic (or a comment) of a clause became a topic (or a comment) in posterior clauses. We found that for all collections, the most frequent relation is TT, then follows CT. TT+CT compose more than 65% of the relationships that we found, whatever the collection is; it is more than 80% for Wikipedia. CC is more rare and TC is the most uncommon relation, especially in Wikipedia.

This preliminary analysis convinced us that using the topic-comment structure could be useful to evaluate readability and that weighting these relations could be a good cue. However, for a scalable method the text structure has to be extracted or annotated automatically. Several parsers have been developed to extract text structure such as HILDA [17] that implements topic changes or SPADE [28] which extracts rhetorical relations and has been used in [22] for example to re-rank documents. These parsers are based on deep analysis of linguistic features and are hardly usable when large volumes of texts are involved. Moreover, they view the topic-comment relation as a remark on the statement while we consider a topic as the phrase that the rest of the clause is understood to be about as in [13].

The information structure is opposed to formal structure of a clause with grammatical elements as constituents. In contrast to a grammatical subject that is a merely grammatical category, a *topic* refers to the information or pragmatic structure of a clause and how it is related to other clauses. However, in a simple English clause, a topic usually coincides with a subject. One of the exceptions are expletives (e.g. *it is raining*) that have only a comment part [15]. Since the unmarked word order in English is *Subject - Verb - Object (SVO)*, we can assume that a topic is usually placed before a verb. As in [13], we also assume that if a subordinate clause provides details on an object, it is rather related to a comment part. Thus, in our method we split a sentence into two parts by a personal verb (not infinitive nor participle) where the first part is considered to be a topic while the rest is viewed as a comment. As opposed to other methods from the literature, this method requires only part-of-speech tagging and its computational complexity is linear over the number of words as well as the number of sentences in a text.

The key idea of our method is that in a coherent text there are relations between topic (or comment) parts of the adjacent sentences and these relations are manifested by word repetition. We represent topic and comment parts of a sentence by bag-of-words. In order to capture the topic-comment relation, we calculate the similarity between them. We propose to use **term** and **noun** based similarities. Since the frequencies of TT, TC, CT and CC differ, it seems reasonable to weight the inter-sentence relationship between topic and comment. Thus, we compute the score between two adjacent sentences s_{i-1} and s_i as the weighted cosine similarity between them:

$$sc(s_{i-1}, s_i) = \frac{1}{||s_{i-1}||||s_i||}[w_{tt}(T_{i-1} \cdot T_i) + w_{ct}(C_{i-1} \cdot T_i) + w_{tc}(T_{i-1} \cdot C_i) + w_{cc}(C_{i-1} \cdot C_i)] \quad (1)$$

where $||\bullet||$ is the length of the corresponding vector, T_i and C_i refer to the bag-of-words representations of topic or comment part of the i-th sentence respectively, the scalar product is marked by $\cdot$, w_{tt}, w_{ct}, w_{tc}, and $w_{cc} \in [0, 1]$ indicate the weights of the TT, TC, CT and CC relations within the text. We estimate text coherence as an average score between adjacent sentences in a text $S = (s_i)_{i=1}^{|S|}$:

$$Coh(S) = \frac{1}{|S|-1} \sum_{i=2}^{|S|} sc(s_{i-1}, s_i) \quad (2)$$

4 EVALUATION

We conducted two series of experiments. For the first evaluation, we used three datasets: (1) **Wikipedia** dump, (2) TREC **Robust**, and (3) **WT10G**. The first dataset is a cleaned English Wikipedia

[1] trec.nist.gov

XML dump of 2012 without notes, history and bibliographic references [3]. We selected 32,211 articles retrieved by the search engine Terrier[2] for the queries from INEX/CLEF Tweet Contextualization Track 2012-2013 [3]. TREC (Text Retrieval Conference) Robust dataset is an unspammed collection of news articles from The Financial Times 1991-1994, Federal Register 1994, Foreign Broadcast Information Service, and The LA Times [31]. We used 193,022 documents retrieved for 249 topics from the Robust dataset. In contrast, WT10G is a snapshot of 1997 of Internet Archive with documents in HTML format, some of which are spam [16]. We retrieved 88,879 documents for 98 topics from TREC Web track 2000-2001. Documents from Robust and WT10G may contain spelling or other errors.

As the first baseline we used a probabilistic graphic model proposed in [18] hereinafter referred to as **Lapata**. Because of page number constraints, we are not detailing this method in this paper. The probabilities were learned from the Wikipedia dataset. For evaluation we calculated text score as the average score between the adjacent sentences. The second baseline **TSP** is a special case of our approach with equal weights for all relations. We also examined a variant of this method where the similarity is based on noun only (**TSPNoun**). We estimated the text coherence as the average cosine similarity between the neighboring sentences.

As in [2, 9, 20, 21, 23], we compare scores assigned to initial documents and the same documents but with randomly permuted sentences. This pairwise evaluation approach was justified in [21]. As in previous approaches, we assumed that the best SO is produced by a human and a good metric should reflect that by assigning higher score to initial SO. Besides, we hypothesized that a good metric has small degradation of results provoked by small permutation in SO and greater rate of shuffling provokes larger effect since the obtained order is remoter from the human-made one. Therefore, as in [12], we consider the following types of datasets: (1) Source collection (**O**), (2) *Rn*-collection (**Rn**), (3) *R*-collection (**R**). *R*-collection is derived from the source collection by shuffling all sentences within each document. *Rn*-collection is generated from the source collection by a random shift of *n* sentences within each document. We used *R*1 and *R*2 collections. The introduction of transitional *Rn*-collections differs from the approaches used in [2, 9, 20, 21, 23].

We calculated system **accuracy** which shows the number of times a system prefers the original order over its permutation divided by the total number of test pairs.

This approach for metric evaluation is completely automatic and requires only a text corpus.

We conducted the second set of experiments on two corpora that are widely used for SO assessment: (1) airplane **Accidents** from the National Transportation Safety Board and (2) articles about **Earthquakes** from the North American News Corpus [9, 21, 23]. Each of these corpora has 100 original texts and for each document 20 permutations (2000 in total). We compared our accuracy results with those reported in the literature, namely entity grid models (**Content + Egrid, Content + HMM-prodn, Content + HMM-d-seq, Egrid + HMM-prodn, Egrid + HMM-d-seq, Egrid + Content + HMM-prodn, Egrid + Content + HMM-d-seq, Egrid + Content + HMM-prodn + HMM-d-seq**)[3], discourse relation based approaches (**Type+Arg+Sal,Arg+Sal, Type+Sal, Type+Arg, Baseline+Type+Arg+Sal**)[4], probabilistic content model (**Probabilistic content**)[5] and topic based model (**Topic-relaxed**)[5].

Table 1: % of times where initial order is scored higher/lower/equally than/to permuted text

Data	Method	O>R1	R1>O	O=R1	O>R2	R2>O	O=R2	O>R	R>O	O=R
Wikipedia	Lapata	38.80	44.07	17.14	38.33	49.26	12.42	30.25	58.96	10.79
	TSP	58.04	26.20	15.75	67.13	25.10	7.77	81.43	13.86	4.71
	TSPNoun	40.64	19.16	40.21	52.96	21.70	25.35	73.17	14.58	12.25
	TopCom	**58.86**	25.99	15.16	**68.12**	24.72	7.16	**83.64**	12.39	3.96
	TCNoun	41.04	19.89	39.08	53.21	22.53	24.25	73.82	14.83	11.35
Robust	Lapata	40.85	50.42	8.73	41.45	55.00	3.55	35.02	63.09	1.89
	TSP	57.15	29.09	13.76	65.85	28.58	5.57	81.77	15.47	2.76
	TSPNoun	44.68	23.67	31.66	55.94	26.87	17.20	75.46	18.56	5.98
	TopCom	**57.66**	29.23	13.11	**66.18**	28.91	4.92	**82.63**	15.45	1.92
	TCNoun	45.14	24.45	30.41	56.07	27.85	16.07	75.57	19.36	5.07
WT10G	Lapata	42.78	51.30	5.92	42.37	55.57	2.06	32.33	66.66	1.01
	TSP	**54.35**	24.02	21.62	**65.42**	24.81	9.78	84.99	12.07	2.95
	TSPNoun	36.22	15.78	48.00	49.00	19.38	31.62	76.69	13.41	9.90
	TopCom	54.31	24.46	21.22	65.24	25.37	9.38	**85.72**	11.69	2.59
	TCNoun	36.42	16.21	47.38	48.91	20.04	31.06	76.84	13.69	9.47

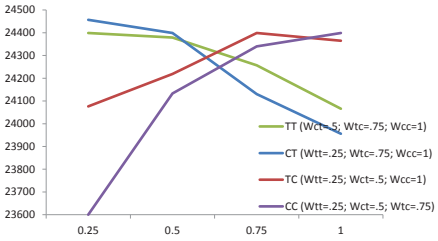

Figure 1: Correlation between w_{tt}, w_{ct}, w_{tc}, w_{cc} & accuracy

In Table 1, *O*, *R*, *R*1 and *R*2 refer to the initial sentence order and the permutations described above and $O > /< /= R\bullet$ shows the proportion of times where initial order was scored higher/lower/equally than/to permuted text for the best set of parameter values $w_{tt} = 0.25$, $w_{ct} = 0.5$, $w_{tc} = 0.75$, and $w_{cc} = 1$. Topic-comment term based method is denoted by **TopCom**. For all collections according to the number of times where the original order was ranked higher than the shuffled one $O > R$, the topic-comment approach outperformed the simple similarity-based metrics and Lapata's baseline. Smaller permutations in sentence order provoke smaller changes in the score. In general noun-based similarity is less accurate than all term based methods. It could be caused by lower probability of non-zero similarity between the adjacent sentences. However, both topic-comment based methods showed better results than their analogues that do not consider text information structure. We varied the coefficients $(w_{tt}, w_{ct}, w_{tc}, w_{cc}) \in \{0.25, 0.5, 0.75, 1\}^4$ on the Wikipedia collection. Figure 1 visualizes the correlation between the number of times where the initial document is preferred to shuffled one $O > R$ and each coefficient with the fixed values of others. Smaller values of w_{tt}, and w_{ct} refer to higher $O > R$, while better results correspond to higher w_{tc} and w_{cc}.

Table 2 presents the results of accuracy on articles about **Earthquakes** and airplane **Accidents** reports. On the **Accidents** dataset

[2] terrier.org is a search engine platform developed by the University of Glasgow

[3] reported as in [23]
[4] reported as in [21]
[5] reported as in [9]

Table 2: Accuracy (%)

Method	Accidents	Earthquakes
TSP	88.7	73.5
TSPNoun	89	60.5
TopCom	86.3	75.1
TCNoun	87	60.4
Content + Egrid	76.8	90.7
Content + HMM-prodn	74.2	95.3
Content + HMM-d-seq	82.1	90.3
Egrid + HMM-prodn	79.6	93.9
Egrid + HMM-d-seq	84.2	91.1
Egrid + Content + HMM-prodn	79.5	95.0
Egrid + Content + HMM-d-seq	84.1	92.3
Egrid + Content + HMM-prodn + HMM-d-seq	83.6	95.7
Probabilistic content	74	-
Topic-relaxed	94	-
Baseline	89.93	83.59
Type+Arg+Sal	89.38	86.50
Arg+Sal	87.06	85.89
Type+Sal	86.05	82.98
Type+Arg	87.87	82.67
Baseline+Type+Arg+Sal	91.64	89.72

we obtained the results comparable with the state of the art. For the **Earthquakes** articles, the accuracy of our system is slightly lower. It can be explained by the following facts: (1) models are trained and tested separately for each dataset [9, 21, 23]; (2) datasets are very homogeneous (some articles are similar up to 90% of words) and, as noted in [9], very constrained in terms of subject and style. In contrast, the coefficients for our method were learned from the Wikipedia collection. This proves that our metric is general and not restricted by a collection but it demonstrates the results comparable with the state of the art machine learning based approaches.

5 CONCLUSIONS

We introduced a novel self-sufficient metric for SO assessment based on topic-comment structure. It has linear complexity and requires only POS-tagging. We evaluated our method on three test collections where it demonstrated high accuracy and significantly outperformed similarity-based baselines as well as a transition probability based approach. The evaluation results allow drawing conclusions that (1) topic-comment methods are more effective than simple similarity based approaches; (2) in general, noun-based similarity is less accurate. In contrast to the state of the art approaches, our method is general and not restricted by a collection but it demonstrates comparable results. One of the promising direction of the future work is the integration of co-reference resolution, synonyms and IDF. Another possible improvement is applying syntactic parsing and linguistic templates for topic-comment structure extraction.

REFERENCES

[1] Regina Barzilay, Noemie Elhadad, and Kathleen R. McKeown. 2002. Inferring Strategies for Sentence Ordering in Multidocument News Summarization. *Journal of Artificial Intelligence Research* (2002), 35–55. 17.

[2] Regina Barzilay and Mirella Lapata. 2008. Modeling local coherence: An entity-based approach. *Computational Linguistics* 34, 1 (2008), 1–34.

[3] Patrice Bellot, Véronique Moriceau, Josiane Mothe, Eric SanJuan, and Xavier Tannier. 2013. Overview of INEX 2013. In *Information Access Evaluation. Multilinguality, Multimodality, and Visualization*. LNCS, Vol. 8138. 269–281. DOI: http://dx.doi.org/10.1007/978-3-642-40802-1_27

[4] Danushka Bollegala, Naoaki Okazaki, and Mitsuru Ishizuka. 2010. A bottom-up approach to sentence ordering for multi-document summarization. *Information processing & management* 46, 1 (2010), 89–109.

[5] Abdelhamid Bouchachia and R Mittermeir. 2003. A neural cascade architecture for document retrieval. In *Proc. of the International Joint Conference on Neural Networks, 2003.*, Vol. 3. IEEE, 1915–1920.

[6] Jieun Chae and Ani Nenkova. 2009. Predicting the fluency of text with shallow structural features: case studies of machine translation and human–written text. *Proc. of the 12th Conference of the European Chapter of the ACL* (2009), 139–147.

[7] J. S. Chall and E. Dale. 1995. *Readability revisited: The new Dale–Chall readability.* MA: Brookline Books, Cambridge.

[8] Kevyn Collins-Thompson and Jamie Callan. 2004. A Language Modeling Approach to Predicting Reading Difficulty. *Proc. of HLT/NAACL* 4 (2004).

[9] Micha Elsner, Joseph L. Austerweil, and Eugene Charniak. 2007. A Unified Local and Global Model for Discourse Coherence.. In *HLT-NAACL*. 436–443.

[10] Micha Elsner and Eugene Charniak. 2008. Coreference-inspired Coherence Modeling. In *Proc. of the 46th Annual Meeting of the ACL on Human Language Technologies: Short Papers (HLT-Short '08)*. ACL, Stroudsburg, PA, USA, 41–44.

[11] Liana Ermakova. 2015. A Method for Short Message Contextualization: Experiments at CLEF/INEX. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction: 6th International Conference of the CLEF Association, CLEF'15, Toulouse, France, September 8-11, 2015, Proceedings*. Springer International Publishing, Cham, 352–363. DOI: http://dx.doi.org/10.1007/978-3-319-24027-5_38

[12] Liana Ermakova. 2016. Automatic Sentence Ordering Assessment Based on Similarity. In *Proc. of EVIA 2016, Tokyo, Japan, 07/06/2016*. NII.

[13] Liana Ermakova and Josiane Mothe. 2016. Document re-ranking based on topic-comment structure. In *X IEEE International Conference RCIS, Grenoble, France, June 1-3, 2016*. 1–10.

[14] E. Fry. 1990. A readability formula for short passages. *Journal of Reading* 8 (1990), 594–597. 33.

[15] Michael Götze, Stephanie Dipper, and Stavros Skopeteas. 2007. *Information Structure in Cross-Linguistic Corpora: Annotation Guidelines for Phonology, Morphology, Syntax, Semantics, and Information Structure*. Interdisciplinary Studies on Information Structure (ISIS), Working papers of the SFB 632, Vol. 7.

[16] David Hawking and Nick Craswell. 2002. Overview of the TREC-2001 web track. *NIST special publication* (2002), 61–67.

[17] Hugo Hernault, Helmut Prendinger, Mitsuru Ishizuka, and others. 2010. HILDA: a discourse parser using support vector machine classification. *Dialogue & Discourse* 1, 3 (2010).

[18] Mirella Lapata. 2003. Probabilistic Text Structuring: Experiments with Sentence Ordering. *Proc. of ACL* (2003), 542–552.

[19] Guy Lebanon and John Lafferty. 2002. Cranking: Combining rankings using conditional probability models on permutations. *Machine Learning: Proc. of the Nineteenth International Conference* (2002), 363–370.

[20] Jiwei Li and Eduard H. Hovy. 2014. A Model of Coherence Based on Distributed Sentence Representation.. In *EMNLP*, Alessandro Moschitti, Bo Pang, and Walter Daelemans (Eds.). ACL, 2039–2048.

[21] Ziheng Lin, Hwee Tou Ng, and Min-Yen Kan. 2011. Automatically Evaluating Text Coherence Using Discourse Relations. In *Proc. of the 49th Annual Meeting of the ACL: Human Language Technologies - vol. 1*. ACL, Stroudsburg, PA, USA, 997–1006.

[22] Christina Lioma, Birger Larsen, and Wei Lu. 2012. Rhetorical Relations for Information Retrieval. In *Proc. of the 35th International ACM SIGIR Conference on Research and Development in Information Retrieval*. 931–940.

[23] Annie Louis and Ani Nenkova. 2012. A Coherence Model Based on Syntactic Patterns. In *Proc. of EMNLP-CoNLL '12*. ACL, Stroudsburg, PA, USA, 1157–1168.

[24] V. Mathesius and J. Vachek. 1975. *A Functional Analysis of Present Day English on a General Linguistic Basis*. Mouton.

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

[26] Eric SanJuan, Véronique Moriceau, Xavier Tannier, Patrice Bellot, and Josiane Mothe. 2012. Overview of the INEX 2011 Question Answering Track (QA@INEX). In *Focused Retrieval of Content and Structure*, Shlomo Geva, Jaap Kamps, and Ralf Schenkel (Eds.). Lecture Notes in Computer Science, Vol. 7424. Springer Berlin Heidelberg, 188–206.

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

[28] Radu Soricut and Daniel Marcu. 2003. Sentence Level Discourse Parsing Using Syntactic and Lexical Information. In *Proc. of NAACL '03 on Human Language Technology - vol. 1*. ACL, 149–156.

[29] AJ Stenner, Ivan Horabin, Dean R Smith, and Malbert Smith. 1988. The lexile framework. *Durham, NC: MetaMetrics* (1988).

[30] Jade Tavernier and Patrice Bellot. 2011. Combining relevance and readability for INEX 2011 Question–Answering track. (2011), 185–195.

[31] Ellen M. Voorhees and Donna Harman. 2000. *Overview of the Sixth Text REtrieval Conference (TREC–6)*.

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

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