Automatic Related Work Section Generation by Sentence Extraction and Reordering

Zekun Deng(✉)[0000−0001−7297−4056], Zixin Zeng, Weiye Gu, Jiawen Ji, and Bolin Hua

Department of Information Management, Peking University, China
dzk@pku.edu.cn

Abstract. Related work section is essential in a scientific publication, for it elaborates past studies relevant to the topic in comparison with the current one. The automatic generation of related work section in scientific papers is a meaningful yet challenging task. While prior works have gained encouraging results, they have not fully addressed the issue of informativeness and the difficulty of obtaining citation sentences due to delay of publication. In this paper, we introduce SERGE, a novel and effective system for generating descriptive related work section automatically by sentence extraction and reordering. Our system first employs a BERT-based ensemble model to select the most salient sentences in reference papers, and then uses a similar model to reorder these sentences for better readability. Automatic evaluation results show that SERGE significantly outperforms existing baselines on ROUGE metrics, gaining an improvement of 18% to 56% on recall and 4% to 33% on F-score. Human evaluation shows that SERGE gains a higher informativeness score than human-written gold standard as well as the baseline, indicating its ability to provide valuable information that matches the real interest of researchers. In contrast to existing methods, since our system is free from delayed citation problem and yields high informativeness, it shows a great potential for various applications.

Keywords: Related work section · Literature review · Scientific documents · Summarization.

1 Introduction

Scientific papers usually contain a related work section, which is also known as a literature review. It summarizes previous works relevant to the research topic in order to establish the link between existing knowledge and new findings[1]. Very often, authors of scientific papers cite existing papers in this section to show the appropriateness of their research question, to justify their adopted methods, and/or to present the creativeness and superiority of their ideas. However, it is quite challenging to produce a high-quality related work section, since it involves identifying crucial points from a long piece of paper and reorganizing them in a neat and logical way.

It is generally accepted that there are two distinct styles of literature reviews: descriptive and integrative[2, 3]. Descriptive literature reviews focus on individual papers and provide more detailed description on the methods, results, and implications of each study. They illustrate previous researches in high accuracy and are thus more objective and rigorous. In contrast, integrative literature reviews focus more on synthesis of ideas. Although including fewer details of individual studies, integrative literature reviews provide more high-level critical summaries of topics and are thus more condensed and structurally complex.[4]

In this paper, we particularly focus on the generation of descriptive related work section. On this matter, Cohan and Goharian[5] have proposed a sentence ranking algorithm that takes advantage of citation context to summarize scientific papers. Abura'ed et al.[6] have proposed a citation-based summarizer for scientific documents based on supervised learning and acheived competitive results in CLScisumm-17 challenges. However, most of the existing studies require citing sentences (a.k.a., "citances") of citing publications as inputs. Thus, these strategies are limited by delay of publications——Mostly a new publication may not be widely recognized and cited within a short period of time and, therefore, it is quite hard to obtain the citing sentences mentioning the publication.

To this end, in this paper, we propose a novel method for automatic generation of descriptive related work section in scientific papers by extracting salient sentences from scientific literature and rearranging them into a logical order. In contrast to most existing methods which suffer from citation delay problem, our method does not require any citances to achieve its goal, making it applicable even when no citation data is available.

The main contributions of this paper are as follows:

1. We propose a novel and effective approach to automatic descriptive related work section generation based on extractive document summarization techniques, including sentence extraction and reordering.
2. Our method does not need any citation data to achieve its goal, which implies that the method does not suffer from the delay of citing publications or require the input of citation data. Such a characteristic offers more potentials of our proposed method with various applications.

2 Related Work

Automatic related work section generation is a special case of multi-document summarization tailored for scientific articles[7]. Multi-document summarization could be either extractive or abstractive, depending on whether the summary contains sentences from source articles[8]. Partly due to scarcity of training data and computational challenges, a large proportion of previous research are in the extractive track, which typically constitutes of a sentence classification subtask and a sentence reordering sub-task[7–9]. Common approaches for extracting relevant sentences include graph-based ranking algorithms[10] and neural classification models[11]. Subsequently, extracted sentences are reordered based on heuristic criteria or neural architectures with sentence ordering mechanisms[12, 13].

Automatic related work generation differs from summarization of generic texts in the following aspects: 1) summarization of generic texts often focus on the content of source papers, whereas the related work section should also delineate contributions and limitations of reference papers (i.e., cited papers) as well as the relationship between reference papers and the current paper; 2) compared to generic texts, scientific articles contain more domain-specific concepts and technical terms, which poses great challenges for language modeling; and 3) scientific articles are more structured than generic texts and reference prior research[8, 14]; accordingly, various unique approaches have been put forward. For instance, Jaidka, Khoo and Na[15] proposed a literature review generation framework that imitates human writing behavior. Many other algorithms are based on citing behavior in scientific articles. Hu and Wan[16] used Probabilistic Latent Semantic Analysis (pLSA) to rank sentences from a set of reference papers. Chen and Zhuge[17] analyzed citation sentences to detect common facts, which were used to find relevant sentences. More recently, Saggoin, Shvets and Bravo[3] exploited pointer-generator architecture with copy-attention technique and coverage mechanism to produce descriptive related work sections.

However, we believe prior research on related work generation, with ROUGE scores as the most popular evaluation metric, haven't fully addressed the issue of informativeness: the property of conveying useful information[9]. This could be potentially problematic because the ROUGE metric may penalize summaries that contain relevant sentences not included in the golden standard summary[8]. Also, as is discussed in Section 1, most of the previous methods are citation-based, which are infeasible in case of delayed citation. Therefore, a novel method is proposed in this paper to tackle these problems.

3 System Design

3.1 Overall Architecture

Here, we introduce our system, **SERGE**, which stands for "**S**entence extraction and r**E**ordering based **R**elated work section **GE**nerator". The overall architecture of SERGE is briefly illustrated in Fig. 1. Given a set of papers from which a related work section will be automatically generated, the system takes the full text of these papers as input, and generates a descriptive related work section involving all these papers as output. SERGE consists of two main parts: a **classification model** and a **reordering model**. The classification model is used to determine whether a sentence is sufficiently salient that it should be included in the generated output. For each sentence as input, the classification model generates a probability value indicating the salience of the sentence. Then, the sentences with highest probability values are fed in to the reordering model, which determines their best order (sequence of sentences). Lastly, the sentences, sorted by the reordering model into the most sensible order, are modified with citation tags and proper pronouns, forming the final output of the system.

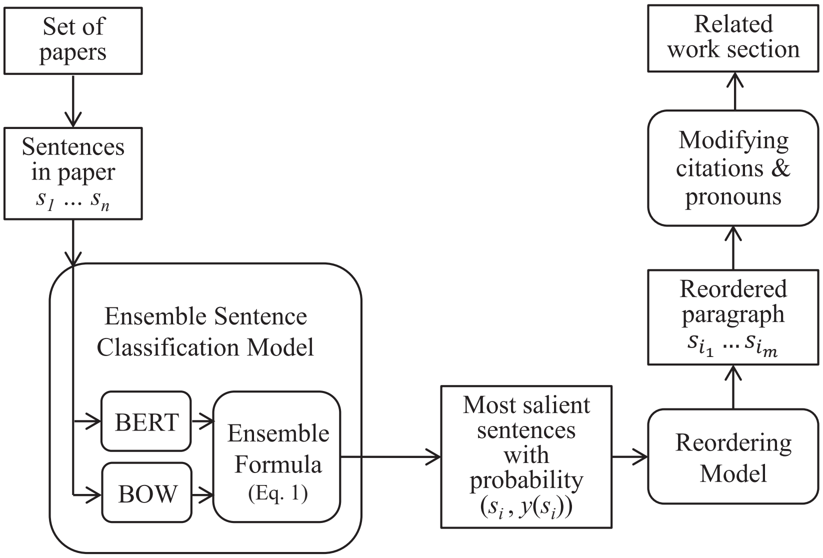
Set of papers
Sentences in paper $s_1 \ldots s_n$
Ensemble Sentence Classification Model
BERT
BOW
Ensemble Formula (Eq. 1)
Most salient sentences with probability $(s_i, y(s_i))$
Reordering Model
Reordered paragraph $s_{i_1} \ldots s_{i_m}$
Modifying citations & pronouns
Related work section

Fig. 1. Overall architecture of SERGE

3.2 Classification Model

The task of classification model is formally stated as follows: Given a sentence s, the model is supposed to assign a label $Y(s) \in \{0, 1\}$ to the sentence, or rather, compute a probability value $y(s) \in [0, 1]$ indicating the salience of sentence s. To accomplish the task, we adopt an ensemble model which consists of two sub-models: a deep neural network model and a bag-of-words (BOW) sentence matching model. The input sentence s is fed in to both models simultaneously. The architecture of the deep neural network model follows Google's original BERT paper[13], as is illustrated in Fig. 2. The input sequence s is first processed by BERT with pre-trained parameters. Then, the final hidden vector of the BERT model $C \in \mathbb{R}^H$ corresponding to the first input token (`[CLS]`) is fed in to a classification layer on the top. The classification layer computes a vector $Z = \text{softmax}(CW^{\mathrm{T}})$, where $W \in \mathbb{R}^{2\times H}$. Here, W is learnable, and $Z = (z_0, z_1)$ is a 2-dimensional vector, where z_i is the estimated probability of the true label of input sequence being i $(i = 0, 1)$.

The training of the model requires annotated data pairs. However, due to the lack of suitable training corpus, we opt to annotate a new dataset automatically by leveraging the ScisummNet corpus[14], a large annotated dataset containing 1000 ACL Anthology papers with their citations. For each paper in the corpus, the dataset includes its full text and incoming citation sentences. Based on the generally agreed assumption that citation sentences usually underscore the most

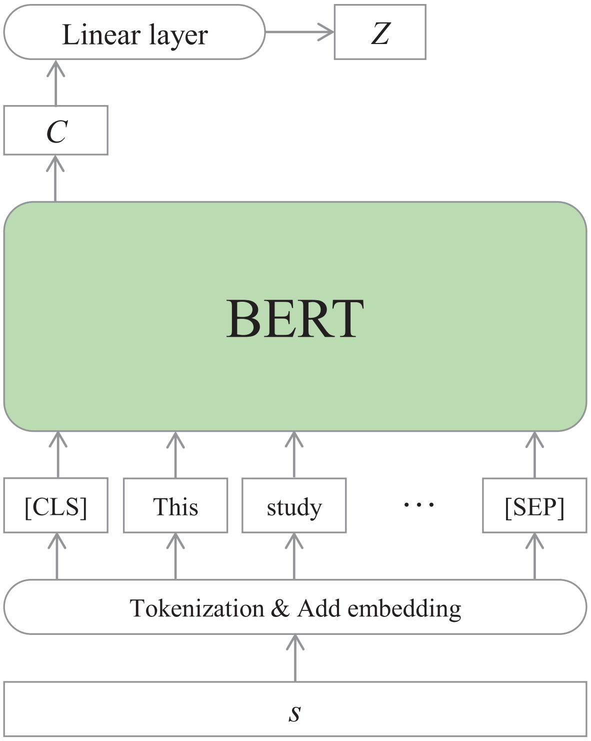

Fig. 2. Architecture of BERT classification model

important aspects of the cited paper and highlight its key contributions, we make use of the citation sentences of a paper in the corpus to produce a gold label of whether a sentence in the paper is salient. An algorithm similar to Nallapati et al.'s paper[11] is applied to annotate the label of each sentence in a paper, which is stated as follows: (1) Join all citation sentences of the paper together to form a benchmark paragraph P^*, and create an empty paragraph P with no sentence in it. (2) Select and append to P the sentence from the abstract or conclusion part of paper which maximizes the ROUGE score between the newly updated paragraph P and P^* and has not been appended to P before. (3) Repeat Step 2 until the ROUGE score does not increase anymore. (4) Label all the sentences included in P as 1 (being salient) and all else as 0 (being not salient).

By employing the greedy annotating algorithm as described above, we obtain an annotated dataset that can be used to train our neural classification model. The dataset includes 11954 training samples, in which 3541 are positive ones.

The bag-of-words sentence matching model simply checks whether the input sentence s contains any of the words in a curated feature word set B. We manually choose the words to be contained in word set B according to the findings of Shin[18], who proposed a dictionary for detecting innovative points in academic literature. Examples of words in B are "novel", "propose", and "improve". Denote the predicted label of this BOW model as $r(s)$, then $r(s) = 1$ if and only if s contains at least one word in B, otherwise $r(s) = 0$.

The combined output of the ensemble model is defined by the following equations:

$$y(s) = \begin{cases} 0, & 0 \leq z_1 \leq \tau_L \\ r(s) \cdot z_1, & \tau_L < z_1 < \tau_H \\ z_1, & \tau_H \leq z_1 \leq 1 \end{cases} \tag{1}$$

$$Y(s) = \begin{cases} 0, & 0 \leq y(s) \leq 0.5 \\ 1, & 0.5 < y(s) \leq 1 \end{cases} \tag{2}$$

Essentially, Eq. 1 considers a trade-off between precision and recall. By setting $\tau_L = 0.2$ and $\tau_H = 0.4$, the ensemble model achieves the most desirable overall performance, with a precision of 0.622 and a recall of 0.793.

3.3 Reordering Model

The task of reordering model is formally stated as follows: Given a set of sentences $S = (s_1, s_2, ..., s_n)$, the model is supposed to find their optimal arrangement $s_{i_1}, s_{i_2}, ..., s_{i_n} (i_r \neq i_t, \forall r \neq t,\ i_k, k, r, t \in \{1, 2, ..., n\})$, where the probability of the sequence $\mathrm{P}(s_{i_1}, s_{i_2}, ..., s_{i_n})$ is maximized. However, considering its sheer scale, it is practically impossible to obtain the solution of the problem directly. Therefore, we decompose the big problem into much smaller ones using a method similar to Chen et al.'s paper[12]. Using the definition of conditional probability, it is obvious that

$$\mathrm{P}(s_{i_1}, s_{i_2}, ..., s_{i_n}) = \mathrm{P}(s_{i_1}) \prod_{k=2}^{n} \mathrm{P}(s_{i_k} | s_{i_1}, s_{i_2}, ..., s_{i_{k-1}}) \tag{3}$$

To simplify the calculation, let

$$\mathrm{P}(s_{i_k} | s_{i_1}, s_{i_2}, ..., s_{i_{k-1}}) = \mathrm{P}(s_{i_k} | s_{i_{k-1}}) \tag{4}$$

where $k \in \{2, 3, ..., n\}$. We also assume $\mathrm{P}(s_{i_1}) = 1$. Thus, Eq. 3 becomes

$$\mathrm{P}(s_{i_1}, s_{i_2}, ..., s_{i_n}) = \prod_{k=2}^{n} \mathrm{P}(s_{i_k} | s_{i_{k-1}}) \tag{5}$$

It can be seen from Eq. 5 that the computation of the probability of all possible arrangements can be approached by simply computing the conditional probability of each sentence appearing after another, reducing the complexity of the problem significantly.

We use a MobileBERT[19] based model to compute the conditional probabilities. MobileBERT is a compact task-agnostic BERT that runs more than 5 times faster than $\text{BERT}_{\text{BASE}}$ while still achieving comparable results on a variety of benchmarks. We adopt MobileBERT instead of BERT mainly for practical reasons: the system running time would get intolerable if BERT is used in real-world application.

To obtain the desired output, MobileBERT is fine-tuned on next sentence prediction (NSP) task, which generates a probabilistic value indicating whether the first sentence in the input is followed by the second one in the source document. The architecture of our NSP model is identical to our neural classification model as is described in Section 3.2, except that the BERT layers are substituted by MobileBERT. The training data for NSP model is also extracted from ScisummNet corpus, whose writing style closely matches the expected input. To build the dataset, we pick out every pair of neighboring sentences from the 1000 papers as positive sample, combined with roughly the same amount of negative samples where the two sentences are not adjacent, yielding a total of 360509 training samples.

Finally, to find the optimal sentence sequence, the model computes the probability of all possible arrangements of the sentences in S using Eq. 5. For the sake of the concision of the final output, if the output of the classification model exceeds $n_{\max} = 3$ sentences, only the ones with highest predicted value $y(s)$ are kept, and the rest are discarded.

4 Evaluation and Results

Both automatic and human evaluation approaches are employed to assess the performance of SERGE. 10 descriptive paragraphs of related work section are randomly extracted from 8 computer science papers published in journals or proceedings. The papers cited in each paragraph are collected to form a set of reference papers. MEAD[20] is used as baseline. During automatic evaluation, SERGE and baseline system generate a related work section respectively on each set of reference papers. The output of the two systems are then compared with the human-written paragraph in published papers by computing the ROUGE score between them.

During human evaluation, 3 computer science experts are instructed to grade the paragraphs generated by the two systems and the gold standard on three aspects: informativeness (INF), fluency (FLU), and succinctness (SUC). The experts are uninformed about the authorship of the texts. The range of score is 1-5.

The result of automatic evaluation is presented in Table 1. Except for the precision of ROUGE-1, SERGE outperforms the baseline on all metrics. Notably,

Table 1. Mean score of automatic evaluation of SERGE and baseline

System	ROUGE-1			ROUGE-2			ROUGE-L		
	P	R	F	P	R	F	P	R	F
SERGE	.258	**.361**	**.297**	**.045**	**.064**	**.052**	**.228**	**.290**	**.254**
MEAD	**.275**	.302	.283	.038	.041	.039	.226	.245	.233

Table 2. Mean score of human evaluation of SERGE, baseline, and gold standard

Source	INF	FLU	SUC
SERGE	**4.23**	3.83	3.83
MEAD	3.97	3.97	3.80
Gold standard	4.10	**4.03**	**4.00**

our system achieves a significant higher recall on all the three ROUGE metrics, yielding a relative gain of 20%, 56% and 18%, respectively. The result of human evaluation is presented in Table 2. SERGE gains an informativeness score of 4.23, which is about 6% higher than the baseline and 3% higher than the gold standard. SERGE also outperforms the baseline on succinctness. In short, the result above indicates that our system performs significantly better than previous baseline in numerous aspects, and even yields a higher informativeness than human-written gold standards.

5 Discussion and Conclusion

In this paper, we propose a novel method for automatic generation of descriptive related work section in scientific papers by extracting salient sentences from past literature and reordering them into a smooth paragraph, which is made possible by two BERT-based neural models. The performance of our method is evaluated by both automatic metrics and human experts. The results show that our method gains a substantial improvement compared with past baselines and achieves a high degree of informativeness comparable to human authors.

Our method addresses the problems of existing works in two ways. First, our method improves the informativeness of automatically generated related work sections significantly, providing more valuable information in existing literature which matches the real interest of researchers. Second, our method is immune from citation delay problem, suggesting its prospect for a wider range of applications.

There are several limitations in the current study. For example, the evaluation is not sufficiently robust due to the high cost of human assessment. Also, our method is not necessarily optimal in fluency and a few other metrics. These issues leave room for future exploration.

The result of this study clearly shows the effectiveness of our novel method for related work section generation. Although the corpus used in this study is limited to the computer science field, it is effortless to adapt our method to other disciplines. Considering its universality and adaptiveness, our method shows a tremendous potential for becoming an intelligent and helpful tool which can increase the efficiency of researchers and boost scientific innovations.

In the future, we may continue to explore new methods for this task via various paths, for instance, by abstractive summarization approaches or entity extraction. We are also interested in integrating summarization problem with certain knowledge bases which brings more intelligence to automatic systems.

Acknowledgements

This work was supported in part by The National Social Science Fundation of China (Number: 17BTQ066).

References

1. Webster, J., Watson, R.T.: Analyzing the past to prepare for the future: Writing a literature review. MIS Quarterly **26**(2), xiii–xxiii (2002), http://www.jstor.org/stable/4132319
2. Jaidka, K., Khoo, C.S., Na, J.C.: Literature review writing: how information is selected and transformed. In: Aslib proceedings: New Information Perspectives. vol. 65, pp. 303–325. Emerald (2013). https://doi.org/10.1108/00012531311330665
3. Saggion, H., Shvets, A., Bravo, À., et al.: Automatic related work section generation: experiments in scientific document abstracting. Scientometrics **125**(3), 3159–3185 (2020). https://doi.org/10.1007/s11192-020-03630-2
4. Khoo, C.S., Na, J.C., Jaidka, K.: Analysis of the macro-level discourse structure of literature. Online Information Review **35**(2), 255–271 (2011). https://doi.org/10.1108/14684521111128032
5. Cohan, A., Goharian, N.: Scientific article summarization using citation-context and article's discourse structure. In: Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing. pp. 390–400 (2015). https://doi.org/10.18653/v1/D15-1045
6. Chiruzzo, L., Saggion, H., Accuosto, P., Bravo, À., et al.: Lastus/taln@ clscisumm-17: Cross-document sentence matching and scientific text summarization systems. In: BIRNDL@ SIGIR (2) (2017)
7. Teslyuk, A.: The concept of system for automated scientific literature reviews generation. In: Krzhizhanovskaya, V.V., Závodszky, G., Lees, M.H., Dongarra, J.J., Sloot, P.M.A., Brissos, S., Teixeira, J. (eds.) Computational Science – ICCS 2020. pp. 437–443. Springer International Publishing, Cham (2020). https://doi.org/10.1007/978-3-030-50420-5_32
8. Ibrahim Altmami, N., El Bachir Menai, M.: Automatic summarization of scientific articles: A survey. Journal of King Saud University - Computer and Information Sciences (2020). https://doi.org/10.1016/j.jksuci.2020.04.020, https://www.sciencedirect.com/science/article/pii/S1319157820303554
9. Liu, Y., Lapata, M.: Hierarchical Transformers for Multi-Document Summarization. arXiv e-prints arXiv:1905.13164 (May 2019)
10. Erkan, G., Radev, D.R.: Lexrank: Graph-based lexical centrality as salience in text summarization. Journal of artificial intelligence research **22**, 457–479 (2004). https://doi.org/10.1613/jair.1523

11. Nallapati, R., Zhai, F., Zhou, B.: SummaRuNNer: A recurrent neural network based sequence model for extractive summarization of documents. Proceedings of the AAAI Conference on Artificial Intelligence **31**(1) (Feb 2017), https://ojs.aaai.org/index.php/AAAI/article/view/10958
12. Chen, X., Qiu, X., Huang, X.: Neural Sentence Ordering. arXiv e-prints arXiv:1607.06952 (Jul 2016)
13. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. arXiv e-prints arXiv:1810.04805 (Oct 2018)
14. Yasunaga, M., Kasai, J., Zhang, R., Fabbri, A., Li, I., Friedman, D., Radev, D.: ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. In: Proceedings of AAAI 2019 (2019). https://doi.org/10.1609/aaai.v33i01.33017386
15. Jaidka, K., Khoo, C., Na, J.C.: Deconstructing human literature reviews – a framework for multi-document summarization. In: Proceedings of the 14th European Workshop on Natural Language Generation. pp. 125–135. Association for Computational Linguistics, Sofia, Bulgaria (Aug 2013), https://www.aclweb.org/anthology/W13-2116
16. Hu, Y., Wan, X.: Automatic generation of related work sections in scientific papers: An optimization approach. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1624–1633. Association for Computational Linguistics, Doha, Qatar (Oct 2014). https://doi.org/10.3115/v1/D14-1170, https://www.aclweb.org/anthology/D14-1170
17. Chen, J., Zhuge, H.: Automatic generation of related work through summarizing citations. Concurrency and Computation: Practice and Experience **31**(3), e4261 (2019). https://doi.org/10.1002/cpe.4261, https://onlinelibrary.wiley.com/doi/abs/10.1002/cpe.4261, e4261 CPE-16-0462.R2
18. Shin, Y.: Research on Innovative Point Identification and Mining of Academic Literature. Master's thesis, Peking University (Jun 2020)
19. Sun, Z., Yu, H., Song, X., Liu, R., Yang, Y., Zhou, D.: MobileBERT: a compact task-agnostic BERT for resource-limited devices. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 2158–2170. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.195, https://www.aclweb.org/anthology/2020.acl-main.195
20. Radev, D., Allison, T., Blair-Goldensohn, S., Blitzer, J., Çelebi, A., Dimitrov, S., Drabek, E.F., Hakim, A., Lam, W., Liu, D., et al.: Mead-a platform for multidocument multilingual text summarization. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation (LREC'04) (2004). https://doi.org/10.7916/D8MG7XZT