Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

ФАКУЛЬТЕТ ИУ «Информатика и системы управления»

КАФЕДРА ИУ7 «Программное обеспечение ЭВМ и информационные технологии»

РАСЧЕТНО-ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К НАУЧНО-ИССЛЕДОВАТЕЛЬСКОЙ РАБОТЕ

НА ТЕМУ:

«Методы выделения составных частей научного текста»

Студент **ИУ7-74Б** _____ **К. А. Рунов**
(Подпись, дата) (И.О.Фамилия)

Руководитель _____ **Ю. В. Строганов**
(Подпись, дата) (И.О.Фамилия)

Консультант _____ **Ю. И. Бутенко**
(Подпись, дата) (И.О.Фамилия)

Рекомендованная руководителем НИР оценка: _____

2024 г.

Министерство науки и высшего образования Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего образования
«Московский государственный технический университет имени Н.Э. Баумана
(национальный исследовательский университет)»
(МГТУ им. Н.Э. Баумана)

УТВЕРЖДАЮ
Заведующий кафедрой ИУ7
_____ И. В. Рудаков
«30» сентября 2024 г.

З А Д А Н И Е
на выполнение научно-исследовательской работы

по теме

Методы выделения составных частей научного текста

Студент группы **ИУ7-74Б**

Рунов Константин Алексеевич

Направленность НИР (учебная, исследовательская, практическая, производственная, др.): учебная.

Источник тематики (кафедра, предприятие, НИР): кафедра.

График выполнения НИР: 25% к 5 нед., 50% к 8 нед., 75% к 11 нед., 100% к 15 нед.

Техническое задание
Провести анализ предметной области технических текстов. Описать известные методы выделения составных частей научного текста. Сформулировать критерии сравнения описанных методов и провести их сравнительный анализ по выбранным критериям.

Оформление научно-исследовательской работы:
Расчетно-пояснительная записка на 12-30 листах формата А4.
Перечень графического (иллюстративного) материала:
Презентация на 3-8 слайдах.

Дата выдачи задания « ___ » сентября 2024 г.

Руководитель НИР _____ **Ю. В. Строганов**

Студент _____ **К. А. Рунов**

Консультант _____ **Ю. И. Бутенко**

РЕФЕРАТ

Отчет 12 с., 3 рис., 0 табл., XXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXXX
источн., YYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYYY
прил.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Структурный анализ документов (Document Layout Analysis, DLA) играет ключевую роль в обработке научно-технических текстов. Такие документы обладают четкой структурой, включающей заголовки, авторов, аннотации, разделы, формулы, таблицы, графики и рисунки [1, 2, 3, 4]. Выявление этих элементов и их логических связей позволяет не только упрощать индексирование и поиск информации, но и улучшать автоматическую обработку текстов, включая аннотирование, реферирование и анализ содержимого.

Документ можно представить в виде иерархии физических модулей (страницы, колонки, абзацы, строки, слова, изображения) или логических модулей (заголовки, авторы, аффилиации, аннотации, разделы, библиография) [5].

Эффективный структурный анализ документов обеспечивает удобную навигацию по тексту, облегчает его разметку и позволяет быстро извлекать необходимые сведения [5].

Целью данной работы является классификация методов выделения составных частей научного текста.

Для достижения поставленной цели необходимо решить следующие задачи:

— провести анализ предметных областей структурного анализа документов и научно-технических текстов;

— провести обзор существующих методов выделения составных частей научного текста;

— сформулировать критерии сравнения описанных методов;

— провести классификацию описанных методов по сформулированным критериям.

1 Анализ предметной области

1.1 Структурный анализ документов

Структурный анализ документов (Document layout analysis, DLA) — процесс сегментирования входного изображения документа на однородные компоненты, такие как блоки текста, рисунки, таблицы, графики и т.д., и их соответствующей классификации [6].

Процесс структурного анализа документов состоит из двух основных этапов — предобработки и анализа структуры документа [2, 5].

На рисунке ниже приведена схема процесса структурного анализа документов.

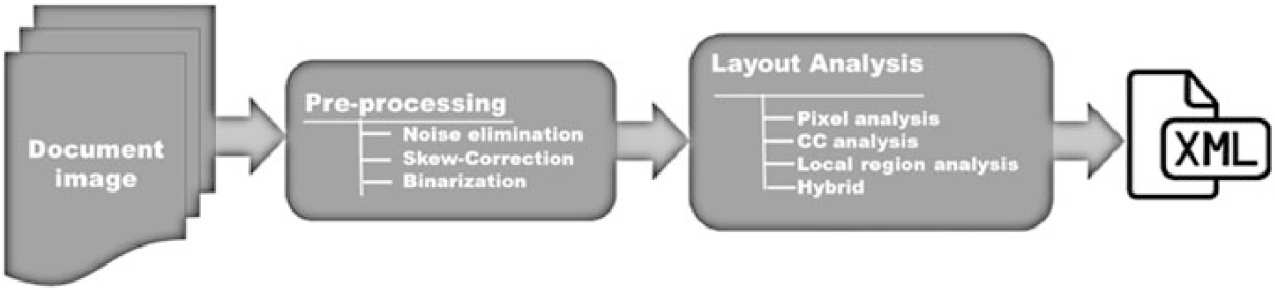

Рисунок 1 – Схема процесса структурного анализа документов [5]

1.1.1 Этап предобработки

1.1.2 Этап анализа структуры документа

1.2 Структура научно-технического текста

[7]

A Cognitive Model for the Representation and Acquisition of Verb Selectional Preferences

Afra Alishahi
Department of Computer Science
University of Toronto
afra@cs.toronto.edu

Suzanne Stevenson
Department of Computer Science
University of Toronto
suzanne@cs.toronto.edu

Abstract

We present a cognitive model of inducing verb selectional preferences from individual verb usages. The selectional preferences for each verb argument are represented as a probability distribution over the set of semantic properties that the argument can possess—a *semantic profile*. The semantic profiles yield verb-specific conceptualizations of the arguments associated with a syntactic position. The proposed model can learn appropriate verb profiles from a small set of noisy training data, and can use them in simulating human plausibility judgments and analyzing implicit object alternation.

1 Introduction

Verbs have preferences for the semantic properties of the arguments filling a particular role. For example, the verb *eat* expects that the object receiving its theme role will have the property of being edible, among others. Learning verb selectional preferences is an important aspect of human language acquisition, and the acquired preferences have been shown to guide children's expectations about missing or upcoming arguments in language comprehension (Nation et al., 2003).

Resnik (1996) introduced a statistical approach to learning and use of verb selectional preferences. In this framework, a semantic class hierarchy for words is used, together with statistical tools, to induce a verb's selectional preferences for a particular argument position in the form of a distribution over all the classes that can occur in that position. Resnik's model was proposed as a model of human learning of selectional preferences that made minimal representational assumptions; it showed how such preferences could be acquired from usage data and an existing conceptual hierarchy. However, his and later computational models (see Section 2) have properties that do not match with certain cognitive plausibility criteria for a child language acquisition model. All these models use the training data in "batch mode", and most of them use information theoretic measures that rely on total counts from a corpus. Therefore, it is not clear how the representation of selectional preferences could be updated incrementally in these models as the person receives more data. Moreover, the assumption that children have access to a full hierarchical representation of semantic classes may be too strict. We propose an alternative view in this paper which is more plausible in the context of child language acquisition.

In previous work (Alishahi and Stevenson, 2005), we have proposed a usage-based computational model of early verb learning that uses Bayesian clustering and prediction to model language acquisition and use. Individual verb usages are incrementally grouped to form emergent classes of linguistic constructions that share semantic and syntactic properties. We have shown that our Bayesian model can incrementally acquire a general conception of the semantic roles of predicates based only on exposure to individual verb usages (Alishahi and Stevenson, 2007). The model forms probabilistic associations between the semantic properties of arguments, their syntactic positions, and the semantic primitives

Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition, pages 41–48,
Prague, Czech Republic, June 2007

Alternating verbs		Non-alternating verbs	
write	0.61	*hang*	0.56
sing	0.67	*wear*	0.71
drink	0.67	*say*	0.75
eat	0.74	*catch*	0.76
play	0.74	*show*	0.77
pour	0.76	*make*	0.78
watch	0.77	*hit*	0.78
pack	0.78	*open*	0.81
steal	0.80	*take*	0.83
push	0.80	*see*	0.87
call	0.80	*like*	0.87
pull	0.80	*get*	0.87
explain	0.81	*find*	0.87
read	0.82	*give*	0.88
hear	0.87	*bring*	0.89
		want	0.89
		put	0.90
Mean:	0.76	Mean:	0.81

Figure 6: Similarity with the base profile for Alternating and Non-alternating verbs.

than verbs with stronger preferences. We use the cosine measure to estimate the similarity between two profiles p and q:

$$\text{cosine}(p,q) = \frac{p \times q}{|p| \times |q|} \quad (9)$$

The similarity values for the Alternating and Non-alternating verbs are shown in Figure 6. The larger values represent more similarity with the base profile, which means a weaker selectional preference. The means for the Alternating and Non-alternating verbs were respectively 0.76 and 0.81, which confirm the hypothesis that verbs participating in implicit object alternations select more strongly for the direct objects than verbs that do not. However, like Resnik (1996), we find that it is not possible to set a threshold that will distinguish the two sets of verbs.

5 Conclusions

We have proposed a cognitively plausible model for learning selectional preferences from instances of verb usage. The model represents verb selectional preferences as a semantic profile, which is a probability distribution over the semantic properties that an argument can take. One of the strengths of our model is the incremental nature of its learning mechanism, in contrast to other approaches which learn selectional preferences in batch mode. Here we have only reported the results for the final stage of learning, but the model allows us to monitor the semantic profiles during the course of learning, and compare it with child data for different age groups, as we do with semantic roles (Alishahi and Stevenson, 2007). We have shown that the model can predict appropriate semantic profiles for a variety of verbs, and use these profiles to simulate human judgments of verb-argument plausibility, using a small and highly noisy set of training data. The model can also use the profiles to measure verb-argument compatibility, which was used in analyzing the implicit object alternation.

References

Abney, S. and Light, M. (1999). Hiding a semantic hierarchy in a Markov model. In *Proc. of the ACL Workshop on Unsupervised Learning in Natural Language Processing*.

Alishahi, A. and Stevenson, S. (2005). A probabilistic model of early argument structure acquisition. In *Proc. of the CogSci 2005*.

Alishahi, A. and Stevenson, S. (2007). A computational usage-based model for learning general properties of semantic roles. In *Proc. of the EuroCogSci 2007*.

Anderson, J. R. (1991). The adaptive nature of human categorization. *Psychological Review*, 98(3):409–429.

Brockmann, C. and Lapata, M. (2003). Evaluating and combining approaches to selectional preference acquisition. In *Proc. of the EACL 2003*.

Ciaramita, M. and Johnson, M. (2000). Explaining away ambiguity: Learning verb selectional preference with Bayesian networks. In *Proc. of the COLING 2000*.

Clark, S. and Weir, D. (2002). Class-based probability estimation using a semantic hierarchy. *Computational Linguistics*, 28(2):187–206.

Collins, M. (1999). *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Holmes, V. M., Stowe, L., and Cupples, L. (1989). Lexical expectations in parsing complement-verb sentences. *Journal of Memory and Language*, 28:668–689.

Levin, B. (1993). *English verb classes and alternations: A preliminary investigation*. The University of Chicago Press.

Li, H. and Abe, N. (1998). Generalizing case frames using a thesaurus and the MDL principle. *Computational Linguistics*, 24(2):217–244.

Light, M. and Greiff, W. (2002). Statistical models for the induction and use of selectional preferences. *Cognitive Science*, 26(3):269–281.

MacWhinney, B. (1995). *The CHILDES project: Tools for analyzing talk*. Lawrence Erlbaum.

Miller, G. (1990). WordNet: An on-line lexical database. *International Journal of Lexicography*, 17(3).

Nation, K., Marshall, C. M., and Altmann, G. T. M. (2003). Investigating individual differences in children's real-time sentence comprehension using language-mediated eye movements. *J. of Experimental Child Psych.*, 86:314–329.

Resnik, P. (1996). Selectional constraints: An information-theoretic model and its computational realization. *Cognition*, 61:127–199.

Рисунок 2


Title
Authors
Abstract
Introduction
Publication
Charts, Graphs, and Equations
Article Text
Conclusion
References

Рисунок 3

2 Формализация задачи

3 Описание существующих методов

3.1 Метод 1

3.1.1 Алгоритм 1

3.1.2 Алгоритм 2

3.2 Метод 2

3.2.1 Алгоритм 1

3.2.2 Алгоритм 2

3.3 Метод 3

3.3.1 Алгоритм 1

3.3.2 Алгоритм 2

4 Классификация существующих методов

ЗАКЛЮЧЕНИЕ

В ходе данной научно-исследовательской работы был проведен анализ предметных областей научно-технических текстов и структурного анализа документов, проведен обзор существующих методов выделения составных частей научного текста, были сформулированы критерии сравнения описанных методов и была проведена классификацию описанных методов по сформулированным критериям.

Таким образом, все задачи для достижения цели данной работы были решены, и цель работы — классификация методов выделения составных частей научного текста — была достигнута.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1. Soto C., Yoo S. Visual Detection with Context for Document Layout Analysis // Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). 2019. С. 3464–3470.

2. Binmakhashen G.M., Mahmoud S.A. Document Layout Analysis: A Comprehensive Survey // ACM Comput. Surv. 2019. Т. 52, № 6.

3. Song M., Rosenfeld A., Kanungo T. Document structure analysis algorithms: A literature survey // Proceedings of SPIE — The International Society for Optical Engineering. 2003. Т. 5010. С. 197–207.

4. Arlazarov et al. Document image analysis and recognition: a survey // Computer Optics. 2022. Т. 46. С. 567–589.

5. Bhowmik S. Document Layout Analysis. — Springer Singapore, 2023 — 86 с.

6. Bhowmik et al. Text and non-text separation in offline document images: a survey // International Journal on Document Analysis and Recognition (IJDAR). 2018. Т. 21.

7. Бутенко Ю.И. Модель текста научно-технической статьи для разметки в корпусе научно-технических текстов // Вестник Новосибирского государственного университета. Серия: Информационные технологии. 2022. Т. 20, № 1. С. 5–13.

ПРИЛОЖЕНИЕ А

Презентация к научно-исследовательской работе содержит XXXXXXXXXXX слайдов.