Research and Exploration on the Path to Intelligent and Knowledge-Based Chinese Semantic Analysis Technology in the News Media Domain: Postprint

Authors: Li Zekui, Sun Fei, Chen Jun

Date: 2023-10-08T00:00:00+00:00

Abstract

Media convergence development is a complex systems engineering endeavor inseparable from the transformation and innovation of technical systems. Against the backdrop of data explosion in the news media domain and the rapid development of artificial intelligence, this paper addresses the various challenges and current status existing in the process of Chinese text semantic analysis within the domestic news media field, and elaborates on and prospects the exploration path toward intelligentization and knowledge-orientation of Chinese text semantic analysis in Xinhua News Agency' s business systems.

Full Text

Abstract

Media convergence development is a complex systematic project that cannot be achieved without transformation and innovation in technological systems. Against the backdrop of explosive growth in news media data and rapid advancement in artificial intelligence, this paper addresses the numerous challenges and current conditions in Chinese text semantic analysis within the domestic news media domain. It elaborates on and prospects the journey toward intelligentization and knowledge-orientation of Chinese text semantic analysis in Xinhua News Agency' s business systems.

Keywords: Chinese semantic analysis; news media domain; intelligent analysis; knowledge analysis

1.1 Intelligent Word Segmentation Combined with News Article Characteristics

Xinhua News Agency processes tens of thousands of articles daily that require text semantic analysis. Behind this wide variety of intelligent analyses lies the fundamental task of word segmentation—the process of dividing text into words in natural language processing. Unlike English text, which uses spaces to separate words, Chinese text requires semantic-based segmentation to recombine continuous characters according to linguistic norms, presenting significantly greater challenges. Focusing on key difficulties in Chinese semantic analysis for the news media domain, such as ambiguity resolution and new word discovery, we have undertaken intelligent exploration from three primary aspects.

1.1.1 Automated Mining of News Media Segmentation Lexicons

Practical segmentation systems often integrate multiple algorithms, but all rely on a high-precision lexicon specific to the news media industry. To address this need and considering the characteristics of our agency' s manuscript texts, we propose a novel word discovery algorithm based on co-occurrence frequency filtering, supplemented by minimal manual verification for dictionary mining. This approach has improved segmentation accuracy to a certain extent.

1.1.2 Building a Comprehensive News Media Domain Corpus

Beyond lexicon-based rule segmentation algorithms, another approach relies on statistical machine learning methods. This methodology depends on a sufficient quantity of "textbooks for machine learning" —training data (corpora) with correctly segmented annotations. To make segmentation models better suited to our agency' s business requirements, we have collected high-quality annotated training datasets from People' s Daily, the National Language Commission, and various evaluation benchmarks, fully leveraging the patterns of Chinese word formation in the news media domain.

1.1.3 Optimization and Enhancement for Entity Phrases

As the national news agency, Xinhua has operated under the direct leadership of the Party Central Committee since its founding, shouldering the sacred mission entrusted by the Party and the people, and serving as the mouthpiece, eyes and ears, think tank, and information hub. Naturally, our manuscripts prioritize correct guidance of public opinion and reflect the mainstream themes of the era. To this end, we have vigorously optimized entity phrases related to current policies and current affairs, such as "One Belt One Road" and "supply-side reform," significantly improving the segmentation capability for relevant phrases. The specific effects are shown in Figure 1 [Figure 1: see original paper].

1.2 Intelligent Topic Classification Based on Knowledge Attributes

Written news reporting represents Xinhua' s traditional and core reporting format, delivering timely, accurate, and authoritative coverage of Party and state policies, as well as important domestic and international news in politics, economics, military affairs, diplomacy, culture, and other fields. To enable intelligent analysis, retrieval, and recommendation of our agency' s written manuscripts, an intelligent topic classification algorithm for news articles is essential.

Currently, Xinhua's knowledge attributes follow a multi-class, multi-level system (13 first-level knowledge attribute categories and over a thousand multi-level index attribute categories). Based on this system, we have established a multi-level topic classification framework (to ensure classification accuracy, it currently reaches a maximum depth of second-level categories, as detailed in Table 1). Combined with popular deep neural network algorithms, we have trained a reliable and efficient intelligent topic classification algorithm.

1.3 Multi-Perspective Intelligent Sentiment Analysis

In major news reporting, Xinhua must not only win the battle for breaking news but also conduct comprehensive, multi-dimensional, and precise statistics and analysis of hot events to maintain correct guidance of public opinion. Sentiment analysis, also known as tendency analysis or opinion mining, serves as a fundamental task in Chinese semantic analysis. News domain sentiment analysis involves analyzing, processing, summarizing, and reasoning about subjective texts with emotional coloring.

Conducting sentiment analysis on hot event news and comments facilitates comprehensive monitoring and management of online public opinion. It enhances capabilities for discovering and handling negative information, intelligence early warning, and public opinion guidance while leveraging internet data to serve the entire news production process. To this end, we propose an algorithm for deep sentiment mining from different perspectives on the same hot event, with each topic' s emotional stance clearly displayed in the interface, as demonstrated in Figure 2 [Figure 2: see original paper].

1.4 Intelligent Automatic Summarization of Text Main Ideas

Automatic text summarization employs intelligent algorithms to automatically compile and generate abstracts. Automatic summarization technology for news texts serves as an auxiliary means to address the current information overload problem in our agency' s vast manuscript materials, helping various users in the "collection, editing, distribution, and supply" workflow to obtain news text information more quickly, accurately, and comprehensively. How to efficiently

store, retrieve, and mine these news texts has become an urgent problem to solve.

For the application scenario of intelligent automatic summarization in the news domain, we have proposed an intelligent automatic summarization method oriented toward news text main ideas. This approach combines knowledge features related to news text structure, syntax, and semantics through extensive iterative optimization and experimentation.

2.1 Knowledge Tag System Combining News Elements and Characteristics

As is well known, the concept of news elements was first proposed by Western journalism: when, where, who, what, why, and how. To bridge news text elements with news knowledge tag extraction, enabling machines to automatically extract news tags more systematically and intelligently, we propose a news tag system comprising five categories: time, location, person, concept, and event. The definitions of concept tags and event tags in this paper are as follows:

Concept Tag: Textual term entities that can be summarized as semantic concepts.
Event Tag: Textual terms that can represent events, directly trigger event generation, and constitute key features determining event categories.

Their classifications and examples are detailed in Table 2 . The news system structure diagram involved in this paper is shown in Figure 3 [Figure 3: see original paper].

2.2 Automatic Knowledge Extraction Based on Tag Categories and Weights

Faced with overwhelming volumes of various types of news and material data, extracting truly useful information represents a threshold for big data applications. Taking our agency' s manuscript texts as an example, when confronted with massive data, this paper first proposes a knowledge tag system specification, then annotates manuscripts according to elements such as time, location, person, concept, and event. The specific algorithm consists of basic Chinese semantic intelligent analysis, phrase merging based on semantic tightness mining, generation and filtering of tag candidate sets, and ranking output based on semantic keyness, as illustrated in Figure 4 [Figure 4: see original paper].

With the establishment of a rich tag system and intelligent extraction algorithm design, many existing problems in Xinhua' s manuscript classification and retrieval will be further alleviated. Simultaneously, we will continue to enhance the system to meet digital network era users' demands for fine-grained search, intelligent retrieval, and personalized customization of manuscripts.

2.3 Preliminary Exploration of Knowledge Graphs for Business Systems

As an important branch of knowledge engineering, knowledge graphs build upon semantic network theory while incorporating excellent algorithms from natural language processing, knowledge representation, and reasoning. They have garnered widespread attention in both industry and academia, driven by big data. The primary purpose of knowledge graph construction is to acquire large-scale, interrelated, computer-understandable knowledge networks.

Since its founding over 80 years ago, Xinhua has accumulated vast amounts of unstructured manuscript texts, semi-structured tables and web pages, and structured data from production systems, containing numerous news knowledge and relationships waiting to be mined (as shown in Figure 5 [Figure 5: see original paper]). This resource resembles a gold mine awaiting development and is extremely valuable.

Knowledge graph construction encompasses many key technologies, ranging from fundamental natural language processing techniques such as precise word segmentation, entity extraction, and syntactic recognition of manuscript texts, to advanced techniques like entity relationship recognition, knowledge fusion, entity linking, and knowledge reasoning.

Given the current situation of scarce domain-specific dictionaries and high costs of manual knowledge annotation, the news domain lacks a standardized, highly usable, and mature knowledge graph construction technology. Addressing these research challenges, various research institutions can complement our agency' s resources to truly develop a knowledge graph technology oriented toward Xinhua' s actual business systems, which we believe will play an important role in solving intelligent analysis problems in news manuscript texts.

Conclusion

This paper introduces research on the journey toward intelligentization and knowledge-orientation of Chinese semantic analysis technology in the news media domain under the major trend of media convergence development. In the intelligent Chinese semantic analysis technology section, we first introduced research on intelligent word segmentation combined with news article characteristics, making segmentation results more aligned with news media business requirements. Second, we briefly explained semantic analysis algorithms from application scenarios, introducing intelligent topic classification, sentiment classification, and automatic summarization technologies.

In the knowledge-based Chinese semantic analysis technology section, we proposed a knowledge tag system combining news elements and characteristics. Based on the actual features of five tag categories, we designed a tag automatic extraction algorithm relying on semantic tightness mining and keyness ranking. Simultaneously, we conducted preliminary exploration and prospects for stan-

dardized, highly usable knowledge graph technology in the news media domain oriented toward Xinhua' s business systems.

References

[1] Zong Chengqing. Statistical Natural Language Processing [M]. Beijing: Tsinghua University Press, 2013.
[2] Li Hang. Statistical Learning Methods [J]. Beijing: Tsinghua University Press, 2012.
[3] Yu Shiwen, et al. Detailed Explanation of Modern Chinese Grammar Information Dictionary [M]. Beijing: Tsinghua University Press, 2003.

(Author Affiliation: Xinhua News Agency Technology Bureau)

Note: Figure translations are in progress. See original paper for figures.