Thematic Structure and Discourse Coherence in Neural Machine Translation of News Discourse: A Comparative Analysis of GPT-4 Based Translate and Google Translate

Shan Wang

School of Foreign Languages and Literature, Beijing Normal University 100875, Beijing, China

Boobi2000@163.com

Abstract. Employing the thematic structure theory, this study investigates the differences in discourse coherence between large language model-based machine translation and traditional neural machine translation in Chinese-English news discourse translation. Findings reveal that large language model-based machine translation more closely resembles human translation in constructing thematic systems and progression patterns, although it may still exhibit limitations in discourse organization compared to human translators. Traditional neural machine translation, on the other hand, tends to overuse constant theme progression, resulting in a lack of discourse hierarchy. This research provides empirical evidence for the application of thematic structure theory in machine translation evaluation and offers insights into optimizing large language model-based machine translation systems to enhance translation coherence.

Keywords: Machine Translation, Large Language Models, Thematic Structure Theory, Discourse Coherence.

1 Introduction

News discourse is an important carrier of international information dissemination and contain rich political, economic and cultural connotations, and are usually written in a formal style, which puts high demands on the coherence of the texts [1]. The development of machine translation technology has provided strong support for the cross-linguistic dissemination of news texts.

In recent years, with the rise of deep learning, the development of machine translation technology provides strong support for the cross-lingual dissemination of news texts. Traditional neural machine translation systems, such as Google Translate, are mainly based on sequence-to-sequence models with Recurrent Neural Network (RNN) or Transformer architectures [2]. Although these systems have achieved some translation quality at the word-sentence level, they still have limitations in terms of discourse coherence, mainly due to the fact that their training data mainly consists of parallel corpora at the sentence level and lacks explicit modeling of discourse-level features [3].

I. A. Khan et al. (eds.), *Proceedings of the 2024 2nd International Conference on Language, Innovative Education and Cultural Communication (CLEC 2024)*, Advances in Social Science, Education and Humanities Research 853,
https://doi.org/10.2991/978-2-38476-263-7_36

Recently, with the rise of deep learning, neural machine translation systems based on large-scale language models have emerged. These systems, such as ChatGPT, utilize language models pre-trained on massive monolingual corpora (e.g., GPT-4), which are adapted to specific translation tasks through techniques such as fine-tuning or cue-engineering, providing powerful semantic representation capabilities for neural machine translation, especially when dealing with discourse coherence [4]. However, although large-scale language models have made promising progress in the field of machine translation, the current research mainly focuses on the translation effect of the models at the word-sentence level [5] [6] [7], and pays insufficient attention to discourse coherence.

To fully unleash the potential of large language models in news discourse translation, it is crucial to analyze their translation characteristics from a discourse perspective. Thematic structure, as an essential means of discourse organization, is closely related to discourse coherence [8]. By comparing and analyzing the realization patterns of thematic structures in translations produced by traditional neural machine translation and large language models, we can reveal their differences in discourse organization and further explore the advantages and limitations of language models in terms of discourse coherence. This is of significant value for further improving the overall performance of neural machine translation in news discourse.

2 Theme-Rheme Structure and Thematic Progression

Thematic theory originates from the functional linguistics research of the Prague School. Mathesius [9] first systematically expounded on the concept of "Theme" in his 1939 article "On So-Called Functional Sentence Perspective". This concept was later expanded by scholars such as Daneš [8] and Firbas [10], gradually developing into a systematic thematic structure theory. Daneš systematically elaborated on the role of thematic structure in discourse organization, proposing three main "thematic progression patterns" while Firbas introduced the concept of "Rheme" as a counterpart to Theme. Halliday & Matthiessen [11] define Theme as "the point of departure of the message" and "the element the speaker selects for 'grounding' what he is going on to say" and Rheme, on the other hand, as "the remainder of the message".

The concept of "thematic progression" was proposed by Daneš [8], referring to the way in which thematic and rhematic information is linked and developed between clauses in a discourse. There are three basic thematic progression patterns: (1) constant theme progression, where the theme of a subsequent clause is related to the theme of the previous clause; (2) simple linear theme progression, where the theme of a subsequent clause is related to the rheme of the previous clause; and (3) split rheme progression, where the rheme of a previous clause is split into multiple items, each becoming the theme in subsequent clauses. McCabe [12] supplemented a fourth pattern based on Daneš' work: split theme, where the theme of a previous clause is split into multiple items, each becoming the theme in subsequent clauses. McCabe also categorized constant theme progression and split theme as thematic progression, and linear theme progression and split rheme as rhematic progression. Building on the original classifications by Daneš [8] and McCabe [12], and with reference to Jiang's [13] research, this study focuses on thematic progression within the scope of the entire text. Thematic progression includes two patterns: constant theme progression (TnTm) and linear theme progression (TnRm). Rhematic progression includes two patterns: constant rhematic progression (RnRm) and linear rhematic progression (RnTm). Additionally, there are two cross-progression patterns, where the theme or rheme of the xth clause is related to the theme plus rheme of a preceding clause (Tx = Tm + Rn and Rx = Tm + Rn). These patterns reveal the logic of information linkage and development in the overall discourse from different levels.

3 Research Methodologies

3.1 Data Collection

The research corpus is selected from bilingual news reports on the China Daily website (http://www.chinadaily.com.cn), which is the most widely circulated English daily newspaper in China with high reputation and guaranteed translation quality. First, we used Python programming to crawl 60 random news articles in both Chinese and English versions from the China Daily website. The English versions serve as human translation corpus (HT), with an average length of 376 words in English and 251 characters in Chinese. Then, the corresponding Chinese news original text was input into Google Translate and OpenAI’s GPT-4 model to obtain traditional neural machine translation corpus (NMT) and large language model-based machine translation corpus (LMT) respectively.

3.2 Research Methods

We take the clause as the basic unit of analysis to examine the similarities and differences in the realization of thematic structure in HT, NMT, and LMT corpora. A clause is the smallest discourse unit that carries thematic information, which can be a main clause, a subordinate clause, or a non-restrictive phrase. For each clause, we manually annotate its theme type according to the classification given by Halliday and Matthiessen [10], which is shown Table 1:

Table 1. Halliday and Matthiessen’s classification of theme types.

Theme type	Description
Topical Theme	Expresses the core experiential content of the clause;
Interpersonal Theme	Realizes interpersonal interaction functions
Textual Theme	Realizes discourse organization functions
Unmarked Theme	Carries the typical topical function of the clause

Theme type	Description
Marked Theme	Generates a marked effect, highlighting the theme or endowing it with special discourse meanings
Simple Theme	Consists of a single topical theme
Multiple Theme	Contains a topical theme accompanied by one or more textual themes and/or interpersonal themes

4 Results and Discussions

4.1 Thematic Complexity

We investigated the realization of single and multiple themes in HT, NMT, and LMT corpora. As demonstrated in Table 2, the three types of corpora exhibit distinct differences in terms of thematic complexity.

Table 2. Overall distribution of simple theme and multiple theme.

Category	HT	NMT	LMT
Simple theme	380 (76.5%)	423 (84.8%)	392 (79.5%)
Multiple theme	117 (23.5%)	76 (15.2%)	101 (20.5%)

From a holistic perspective, single themes predominate in all three corpora, which aligns with the stylistic characteristics of news discourse that objectively present facts and focus on core information. Among the corpora, the NMT corpus has the highest proportion of single themes (84.8%), followed by LMT (79.5%) and HT (76.5%); the distribution of multiple themes exhibits the opposite trend. This indicates that, compared to human translation, machine translation is more inclined to employ simple theme-rheme structures, with the NMT model being particularly prone to this tendency. This may be attributed to NMT's emphasis on local translation correspondence and its lack of overall discourse control. In contrast, the LMT model, trained on massive corpora, has acquired, to a certain extent, the general patterns of English discourse organization.

We further examined the internal composition of multiple themes. As shown in Table 3, there exist differences among HT, NMT, and LMT in terms of the distribution of various types of multiple themes.

Table 3. Overall distribution of the sub-types of multiple theme.

Category	HT	NMT	LMT
Topical+Textual	84 (71.8%)	56 (73.7%)	76 (75.2%)
Topical+interpersonal	18 (15.4%)	8 (10.5%)	12 (11.9%)
Topical+Textual+ Interpersonal	15 (12.8%)	12 (15.8%)	13 (12.9%)

Specifically, the "textual + topical" type is the most frequently occurring multiple theme in all three corpora (>70%), which aligns with the requirement for logical cohesion in news discourse. The proportion of "Topical+ Interpersonal" themes is higher in the HT corpus compared to NMT and LMT, possibly due to human translators' superior ability to construct interpersonal meaning through modal markers, evaluative components, and other means, thereby enhancing interaction with readers. In contrast, NMT and LMT exhibit a relative lack of interpersonal themes, indicating a need for strengthening discourse interactivity. Moreover, complex themes containing textual, interpersonal, and topical components appear most frequently in the NMT corpus, which may be related to NMT's "sentence-by-sentence correspondence" strategy with respect to the source language form, leading to redundant and cumbersome theme structures. The usage of triple themes in HT and LMT corpora is relatively similar, suggesting that LMT has, to a certain extent, imitated the discourse organization characteristics of human translation.

We further compared the high-frequency theme forms across the three corpora. The HT corpus features more complex textual theme words (e.g., "meanwhile" "furthermore" "however") that introduce diverse logical-semantic relations, while the NMT corpus primarily uses basic conjunctions (e.g., "and" "but"). The LMT corpus, with words like "additionally" and "nevertheless" lies in between but still lacks the richness of human translation, indicating an expanded diversity of cohesion devices compared to NMT, albeit with room for improvement in flexibility. Regarding interpersonal themes, the HT corpus frequently uses modal words (e.g., "may" "would" "definitely"), reflecting a stronger tendency towards subjective interpretation. The NMT corpus mostly employs auxiliary verbs in interrogative sentences, exhibiting weaker subjectivity and interactivity. While the LMT corpus contains some evaluative adverbs (e.g., "obviously" "probably" "actually"), it still falls short of the richness and appropriateness of human translation.

4.2 Theme-Rheme Progression

Based on the exploration of local theme selection, this section examines the thematic progression patterns at the discourse level in HT, NMT, and LMT corpora. Following the classification of thematic progression proposed by Daneš (1974) and others, we calculated the distribution of six progression patterns (TT, TR, RR, RT, T=T+R, R=T+R) in the three corpora.

Table 4 presents the raw frequencies and percentages of the six thematic progression patterns in the HT, NMT, and LMT corpora. It is evident that TT and TR are the most frequently occurring progression methods in all corpora, with their combined proportions exceeding 70%. This indicates that advancing discourse through themes as a cohesive device is a primary characteristic of English news discourse. It is noteworthy that the proportion of the TT pattern in the NMT corpus is as high as 63.2%, higher than that in HT (49.4%) and LMT (51.9%). In contrast, NMT's use of the TR pattern is markedly insufficient (16.6%), far lower than that in HT (32.5%) and LMT (26.4%).

Table 4. Overall distribution of the theme-rheme progression.

Category	HT	NMT	LMT
TnTm	246 (49.4%)	312 (63.2%)	268 (51.9%)
TnRm	162 (32.5%)	82 (16.6%)	136 (26.4%)
RnRm	30 (6.0%)	42 (8.5%)	53 (10.3%)
RnTm	34 (6.8%)	26 (5.3%)	19 (3.7%)
Tx = Tm + Rn	13 (2.6%)	14 (2.8%)	19 (3.7%)
Rx = Tm + Rn	13 (2.6%)	18 (3.6%)	21 (4.1%)

There could be several reasons for this phenomenon: First, TT is conducive to maintaining discourse continuity and topic consistency, aligning with the objective reporting characteristic of news discourse, and thus is widely used in all corpora. Second, TR introduces new topics by carrying forward the rheme information from the previous discourse segment, promoting discourse development while maintaining coherence, and is a common cohesive strategy in English discourse [14]. In comparison, NMT relies more on TT, showing a slight lack of extensibility in discourse, while LMT compensates for this limitation to a certain extent by using more theme-rheme alternation patterns to introduce new information, making the discourse more hierarchical. This may be related to LMT's greater emphasis on integrating contextual information compared to NMT. Finally, the frequency of cross-progression patterns (Tx = Tm + Rn, Rx = Tm + Rn) is relatively low in all corpora, indicating that the method of introducing topics by combining multiple discourse segments is less common, which may be related to the hypotactic characteristics of English, i.e., the use of complex grammatical structures is less frequent.

(1) China has committed to reaching peak carbon emissions before 2030 and achieving carbon neutrality by 2060. This commitment has helped to mobilize state institutions, boost the confidence of businesses, raise public expectations, and serve as a critical policy decision for expanding and accelerating green development efforts. (HT)

(2) China has proposed the goal of striving to peak carbon dioxide emissions before 2030 and strive to achieve carbon neutrality before 2060. The “double carbon” goal will help mobilize government agencies, enhance business confidence, raise public expectations, and become a key decision to promote green development. (NMT)

(3) China has committed to reaching peak carbon emissions before 2030 and achieving carbon neutrality by 2060. This dual-carbon commitment has helped mobilize government agencies, boost business confidence, raise public expectations, and serves as a critical policy driving force for promoting green development. (LMT)

To further illustrate the differences in thematic progression patterns across the three corpora, we selected three typical examples for analysis. As shown in example (1), the first sentence in the HT corpus introduces the theme “China has committed to...”, which serves as the topical theme (T). The second sentence takes “This commitment” as the theme, carrying forward the rheme (R) of the previous sentence. This is a typical TR, demonstrating thematic coherence and progression. In the NMT corpus, the first sentence introduces “China has proposed the goal...” as the topical theme (T), and the theme of the second sentence, “The ‘double carbon’ goal”, forms a TT pattern with it. Although the rheme parts of the two sentences vary, the repetition of the theme results in a lack of thematic extension and the introduction of new information. In the LMT corpus, the theme of the first sentence is the same as in HT, and the second sentence takes “This dual-carbon commitment” as the theme, referring back to the rheme of the previous sentence while also introducing a new rheme, “has helped mobilize...”. This similarly reflects the TR pattern, maintaining coherence while advancing the development of the discourse. However, LMT directly repeats “dual-carbon commitment” in its expression, which appears somewhat redundant and less concise compared to HT’s use of “This commitment”.

5 Conclusions

This study investigates the usage of thematic systems and thematic progression patterns in machine translation news discourse. Overall, the results shows that the human translation showcases the most dynamic use of simple linear theme progression, while the traditional neural machine translation model overuses constant theme progression, compromising coherence and hierarchy. The language model-based machine translation lies in between, better integrating context and introducing new topics than NMT, but still lacking the flexibility of human translation. Findings suggest that incorporating flexible human translation strategies and thematic progression patterns is crucial for advancing machine translation systems from mere “translators” to “discourse experts”.

References

1. Bielsa, E., Bassnett, S.: Translation in global news. Routledge (2008).
2. Bahdanau, D., Cho, K., Bengio, Y.: Neural machine translation by jointly learning to align and translate. In: Bengio, Y., LeCun, Y. (eds.) 3rd International Conference on Learning Representations, Conference Track Proceedings (2015).
3. Läubli, S., Sennrich, R., Volk, M.: Has machine translation achieved human parity? A case for document-level evaluation. arXiv preprint arXiv:1808.07048 (2018).
4. Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics 8, 726-742 (2020).
5. Conneau, A., Lample, G.: Cross-lingual language model pretraining. In: Advances in Neural Information Processing Systems, pp. 7059-7069 (2019).
6. Imamura, K., Sumita, E.: Recycling a pre-trained BERT encoder for neural machine translation. In: Proceedings of the 3rd Workshop on Neural Generation and Translation, pp. 23-31 (2019).
7. Weng, R., Yu, H., Huang, S., Cheng, S., Luo, W.: Acquiring knowledge from pre-trained model to neural machine translation. In: Proceedings of the AAAI Conference on Artificial Intelligence, vol. 34, no. 05, pp. 9266-9273 (2020).
8. Daneš, F.: Functional sentence perspective and the organization of the text. In: Daneš, F. (ed.) Papers on functional sentence perspective, pp. 106-128. Academia, Prague (1974).
9. Mathesius, V.: O takzvaném aktuálním cleneni vety (On the so-called functional sentence perspective). In: Kuno, S. (ed.) Harvard Studies in Syntax and Semantics, pp. 467-480 (1975).
10. Firbas, J.: Functional sentence perspective in written and spoken communication. Cambridge University Press (1992).

11. Halliday, M.A.K., Matthiessen, C.M.I.M.: Halliday's introduction to functional grammar, 4th edn. Routledge (2014).
12. McCabe, A.: Theme and thematic patterns in Spanish and English history texts. Ph.D. dissertation, Aston University (1999).
13. Jiang, Y., Niu, J.: How are neural machine-translated Chinese-to-English short stories constructed and cohered? An exploratory study based on theme-rheme structure. Lingua 273, 103318 (2022).Liu, Y., Gu, J., Goyal, N., Li, X., Edunov, S., Ghazvininejad, M., Lewis, M., Zettlemoyer, L.: Multilingual denoising pre-training for neural machine translation. Transactions of the Association for Computational Linguistics 8, 726-742 (2020).
14. Quirk, R., Greenbaum, S., Leech, G., Svartvik, J.: A comprehensive grammar of the English language. Longman, London (1985).