AlignAR: Generative Sentence Alignment for Arabic-English Parallel Corpora of Legal and Literary Texts

Baorong Huang, Ali Asiri*
School of Foreign Languages, Huaihua University
Alith University College, Umm al-Qura University
Huaihua, China,
Makkah, Saudi Arabia
huangbaorong2021@gmail.com, amaasiri@uqu.edu.sa

Abstract

High-quality parallel corpora serve as the fundamental backbone for advancements in Machine Translation (MT) research and the development of effective translation pedagogy. Despite this need, robust resources for the Arabic-English language pair remain significantly scarce. Furthermore, existing datasets are often limited by their reliance on simplistic one-to-one sentence mappings, which fail to capture the structural complexities inherent in natural language translation. To address this deficiency, this paper presents AlignAR, a novel generative sentence alignment method, alongside a comprehensive new Arabic–English dataset that juxtaposes simple legal documents with complex literary texts. Our evaluation demonstrates that "Easy" datasets lack the discriminatory power to fully assess alignment methods. By reducing one-to-one mappings within our "Hard" subset, we exposed the limitations of traditional alignment techniques when faced with structural divergence. In contrast, Large Language Model (LLM) based approaches demonstrated superior robustness and adaptability. Specifically, the proposed LLM-based approaches demonstrated better robustness, achieving an overall F1-score of 85.5%, a nearly 9% improvement over previous methods. This study underscores the importance of complex benchmarks and validates the efficacy of generative models in handling the intricacies of bitext alignment. The codes and datasets are available on GitHub.

Keywords: Parallel Corpora, Generative Sentence Alignment, Arabic–English Translation

1. Introduction

Parallel sentence alignment is an important step in preparing parallel data for various Natural Language Processing (NLP) applications, particularly machine translation. It involves identifying corresponding sentences between two texts that are mutual translations, as shown in Figure 1. Given that manual alignment is resource-intensive and time-consuming, various automatic methods are proposed to address the challenges.

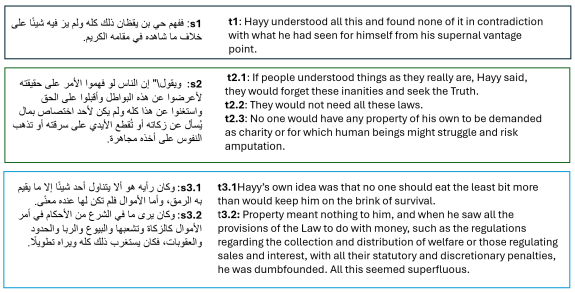

Figure 1: Examples of 1-to-1, 1-to-many and many-to-many alignments between source and target sentences written in Arabic and English, respectively.

Over the past three decades, the research has progressed from early statistical and length-based models (Brown et al., 1991; Gale and Church, 1993) to heuristic, structural, and modern neural embedding-based approaches capable of handling diverse domains, combining multilingual sentence embeddings, anchor-point detection, and dynamic alignment algorithms, such as dictionary-based extraction(Althobaiti, 2022), divide and conquer(Steingrímsson et al., 2023; Zhang, 2022). More recent work explores embedding-based methods, such as overlapping fixed-length segmentation(Wang et al., 2024), multilingual embeddings (Thompson and Koehn, 2019; Liu and Zhu, 2023), contextualized document-aware sentence embeddings(Molfese et al., 2024), and particle swarm optimization(Shang and Li, 2024). Despite these advances, research on sentence alignment using Large Language Models (LLMs) remains notably scarce.

Previous alignment methods are evaluated on web datasets in English-German texts(Koehn et al., 2018) or some low-resource languages, such as Pashto–English(Koehn et al., 2020) or Estonian-Lithuanian(Sloto et al., 2023). In addition, some gold alignment datasets are released to evaluate the alignment methods for French-German pairs such as Text+Berg, Chinese-English pairs(Liu and Zhu, 2023), English-Icelandic pairs(Steingrímsson et al.,

Corresponding author: amaasiri@uqu.edu.sa

Proceedings of The 7th Workshop on Open-Source Arabic Corpora and Processing Tools (OSACT7) with 5 Shared Tasks, pages 59–65
May 11, 2026.

2023), or various European languages(Molfese et al., 2024). However, the resources on English-Arabic pairs are underexplored.

To address these limitations, we conduct an in-depth investigation into the integration of Large Language Models (LLMs) for automated sentence alignment and put forward the following two main contributions:

- We introduce a novel LLM-based alignment method. This method leverages the contextual understanding of LLMs to generate high-precision initial alignments, which we rigorously evaluate against our specialized parallel corpus;
- We present the Arabic–English Parallel Corpus, a new gold-standard dataset designed to test alignment robustness across varying levels of difficulty. The corpus is categorized into a legal sub-corpus, representing structured technical prose, and a literary sub-corpus, containing complex, non-linear narratives that pose a significant challenge to previous alignment methods.

Ultimately, we hope that this study provides both a scalable methodology and a high-fidelity resource to stimulate further research into specialized bilingual alignment, particularly for low-resource and linguistically complex pairs.

2. Existing methods

2.1. BleuAlign

Sennrich and Volk (2010) proposed Bleualign, in which high-scoring sentence pairs are treated as anchor points, while dynamic programming is employed to infer optimal alignments between anchors, allowing for one-to-one and one-to-many mappings. The approach also integrates traditional length-based heuristics as a fallback mechanism in difficult cases.

2.2. VecAlign

Thompson and Koehn (2019) introduced Vecalign, which represents sentences from both languages in a shared semantic vector space using pre-trained multilingual encoders and computes cosine similarity to identify semantically corresponding sentence pairs. To ensure scalability, the authors proposed a linear-time dynamic programming algorithm that aligns sentences monotonically while permitting sentence splits and merges.

2.3. BertAlign

Liu and Zhu (2023) proposed Bertalign, in which sentences are encoded using transformer embeddings, enabling more precise similarity estimation than static word or sentence embeddings. Bertalign computes sentence-level similarity scores and applies dynamic programming to determine optimal alignments under monotonicity constraints.

3. Datasets

The data used in this study are divided into two main categories based on the complexity of the sentence alignment task: easy-alignment data and hard-alignment data, as shown in Table 1. the hard part consists of 5 documents, totaling 378 Arabic sentences and 774 English sentences, with the smallest source/target ratio around 0.45, which means that one Arabic sentence aligns to at least two English sentences on average. The easy part also consists of 5 documents manually aligned from Arabic laws and its English translations, totaling 892 source lines and 1093 target lines, with the smallest source/target ratio around 0.74, much higher than the hard part.

The easy-alignment subset comprises modern legal texts characterized by high structural parallelism and terminological consistency. These data were collected from the Saudi National Center for Archives and Records (NCAR), which provides official Saudi laws alongside their English translations, including the Companies Law, Anti-Bribery Law, and Anti-Commercial Concealment Law. Due to their formulaic drafting style, controlled vocabulary, and close translation correspondence, these legal texts represent low-complexity alignment data.

In contrast, the data of the hard-alignment subset are drawn from classical and literary Arabic texts, which present significant challenges due to archaic language, stylistic density, and non-literal translation strategies. The raw data are from the open-source platform Rasaifa, including the philosophical narrative Ḥayyibn Yaqẓān, and short stories included in Hassan Ghazala's A Textbook of Literary Translation (2013).

4. Method

In this study, we frame document alignment as a zero-shot inference task, prompting the models to identify correspondences based on semantic equivalence rather than lexical overlap. We propose a novel sentence alignment that considers alignment as a translation mapping task, and request the LLM to identify the translations of the source text, rather than directly performing the alignment.

Table 1: Statistics of the Arabic–English Parallel Corpus

Dataset	#AR-SENT	#AR-TKN	#EN-SENT	#EN-TKN	SENT. (%)	1-1 (%)
Easy 1	153	4,405	206	7,783	74.27	78.15
Easy 2	202	6,437	257	11,438	78.60	77.78
Easy 3	161	4,726	198	8,471	81.31	81.76
Easy 4	202	4,893	233	8,766	87.73	85.20
Easy 5	174	5,770	199	10,650	87.44	87.21
Hard 1	93	3,880	202	6,318	46.03	28.77
Hard 2	100	3,998	194	6,823	51.55	40.26
Hard 3	93	3,111	175	5,849	53.14	45.78
Hard 4	92	3,068	203	5,800	45.32	40.70
Hard 5	101	2,759	223	6,906	45.29	31.18

Note: #AR-SENT and #AR-TKN indicate sentence count and token count in the Arabic text; #EN-SENT and #EN-TKN indicate sentence count and token count in the English text; SENT(%) means the ratio of Arabic sentences to English sentences in the dataset; 1-1 (%) means the frequency (percentage) of 1-to-1 alignments in each dataset. Both source and target texts have been tokenized with Stanza before counting tokens.

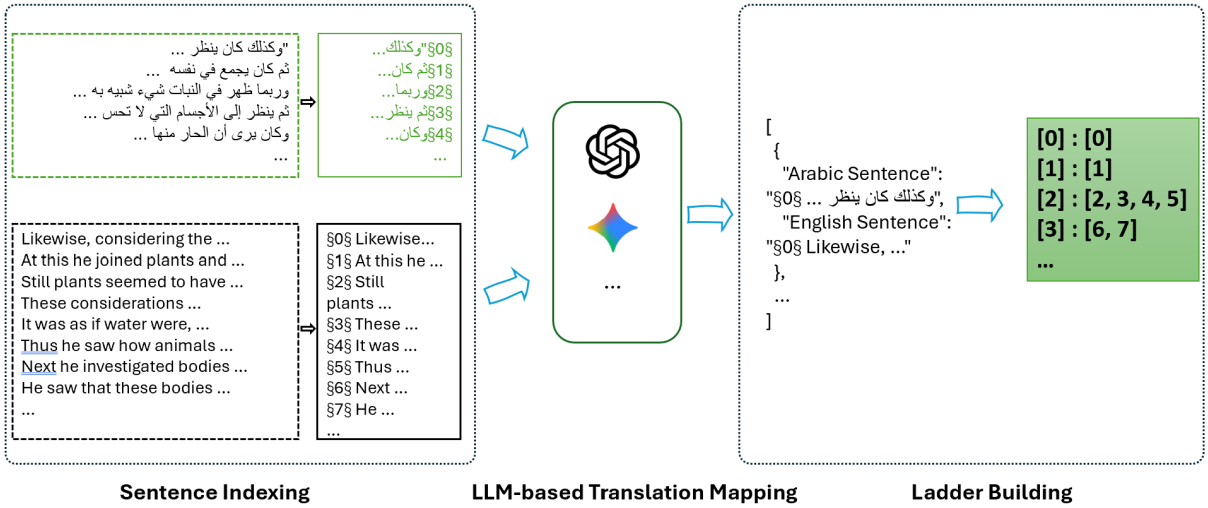

Figure 2: Our alignment procedure consists of three steps: (1) sentence indexing, i.e., explicitly indexing the lines with sequence number; (2) LLM-based translation mapping, i.e., requesting the selected LLM to select the translations from the target lines; (3) ladder building, extracting the index number from the formatted JSON output and obtaining ladder files for evaluation.

4.1. Alignment Procedure

We propose a three-step alignment procedure. In the first step, we add a sentence index (sequence number) as $[0\ s_0,\ \ldots,\ n\ s_m]$ and $[0\ t_0,\ \ldots,\ n\ t_n]$ to each source line and each target line. Furthermore, to further reduce the inadvertent addition or deletion of words by an LLM, we change the parallel aligning task into a translation mapping task, requesting the LLM to pick out the corresponding translations of the source lines from the target lines and output the entire mappings in a JSON format, together with their sentence indexes. This design addresses the issue that directing the LLM to output only sentence indices considerably degrades alignment quality. To mitigate this, we require the LLM to reproduce the original text alongside the indices. In the final step, we only extract the sentence indexes to obtain a ladder file that only contains line number mappings for evaluation, as shown in Figure 2. In this way, we obtained the precise sentence indexes required for alignment, while avoiding the issues in mapping the texts generated by LLM to the original texts, which is complicated because sometimes LLM failed to generate the texts identical with the input. For the prompt used, please see Appendix 10.1.

4.2. Evaluation Metrics

We use the strict precision, strict recall and strict F1 as our evaluation metrics. Strict true positive, as

Table 2: Comparison across different methods on easy document alignment

Item	Metric	Baselines			Ours	
		BERT	BLEU	VEC	GPT	Gemini
Easy 1	P	0.936	0.966	0.961	**0.987**	**0.987**
	R	0.961	0.947	0.961	**0.993**	**0.993**
	F_1	0.948	0.957	0.961	**0.990**	**0.990**
Easy 2	P	0.985	0.919	0.985	**0.990**	0.980
	R	0.980	0.914	0.985	**0.995**	0.990
	F_1	0.982	0.916	0.985	**0.993**	0.985
Easy 3	P	0.981	0.925	0.975	**1.000**	**1.000**
	R	0.981	0.925	0.956	**1.000**	**1.000**
	F_1	0.981	0.925	0.965	**1.000**	**1.000**
Easy 4	P	**1.000**	0.944	0.970	0.980	0.990
	R	**1.000**	0.954	0.965	0.990	0.995
	F_1	**1.000**	0.949	0.967	0.985	0.992
Easy 5	P	**1.000**	0.965	0.994	0.989	**1.000**
	R	**1.000**	0.965	0.988	0.994	**1.000**
	F_1	**1.000**	0.965	0.991	0.991	**1.000**
Overall	P	0.982	0.943	0.977	0.989	**0.991**
	R	0.985	0.943	0.972	0.994	**0.995**
	F_1	0.984	0.942	0.974	0.992	**0.993**

Note:BERT means BertAlign, BLEU means BlueAlign and VEC means VecAlign. Bold indicates the best-performing method.

defined by Sennrich and Volk (2010), means the alignment hypothesis should be identical to the reference alignment. We adopt the strict metrics because the actual usage of the alignments in translation studies or MT training requires the precise and exact mappings between the source lines and target lines. In this case, the overlapping alignment as lax measures allowed will be considered incorrect. We take the Micro-Averaged Metrics for computing the overall scores for the easy and hard parts.

5. Experiments & Results

5.1. Experimental settings

We evaluate the alignment accuracy on the Arabic-English datasets using our proposed LLM-based alignment methods, in comparison with the previous methods, including BertAlign, BleuAlign, and VecAlign. Our experiments are conducted on a computer with Intel Ultra 7 265KF, 64GB RAM, and NVIDIA RTX 5070 Ti. Two LLMs are used in the experiments, including GPT-5.1-mini and Gemini-2.5-flash.

5.2. Experiments on legal text (easy part)

An analysis of the experimental results in Table 2 suggests that the "Easy" subsets of the dataset may lack the discriminatory power required to fully evaluate the capabilities of aligning methods. As shown by the "Overall" metrics, Gemini achieves a near-perfect F_1 score of 0.993; however, the consistently high scores across all methodologies, all exceeding the 0.90 threshold, indicate that these specific benchmarks have reached saturation. Consequently, more challenging data are required to reveal the latent performance gaps.

Table 3: Comparison across different methods on hard document alignment

Item	Metric	Baselines			Ours	
		BERT	BLEU	VEC	GPT	Gemini
Hard 1	P	0.667	0.507	0.670	0.684	**0.811**
	R	0.757	0.479	0.735	0.783	**0.880**
	F_1	0.726	0.493	0.701	0.730	**0.844**
Hard 2	P	0.724	0.563	0.791	0.718	**0.871**
	R	0.807	0.519	0.818	0.841	**0.920**
	F_1	0.763	0.540	0.804	0.775	**0.895**
Hard 3	P	0.606	0.410	0.645	0.612	**0.688**
	R	0.648	0.410	0.682	0.682	**0.750**
	F_1	0.626	0.410	0.663	0.645	**0.717**
Hard 4	P	0.778	0.618	0.826	0.884	**0.903**
	R	0.865	0.547	0.854	**0.944**	**0.944**
	F_1	0.819	0.580	0.840	0.913	**0.923**
Hard 5	P	0.699	0.340	0.780	0.748	**0.880**
	R	0.742	0.310	0.804	0.794	**0.907**
	F_1	0.720	0.324	0.792	0.770	**0.893**
Overall	P	0.696	0.494	0.743	0.729	**0.831**
	R	0.757	0.462	0.780	0.809	**0.831**
	F_1	0.726	0.477	0.761	0.767	**0.855**

5.3. Experiments on literary text (hard part)

The results on the challenging document alignment task (Table 3) provide a much clearer differentiation of model capabilities. As the task complexity increases, the "ceiling effect" observed in the easy datasets disappears, revealing significant discrepancies in model robustness. Gemini achieves an overall F_1 score of 0.855, which represents a substantial margin over the BertAlign baseline (0.726) and the GPT-based approach (0.767). The sharp decline in performance across all baseline methods in "Hard" datasets underscores their limitations. For instance, BleuAlign's overall F_1 plummeted from 0.942 to 0.477, and BertAlign's overall F_1 dropped from 0.984 to 0.726. More importantly, the divergence between GPT and Gemini on the hard datasets—most notably in Hard 5, where Gemini outperforms GPT by over 12 percentage points in F_1—indicates that the selection of LLMs also exerts considerable influence on the alignment quality.

6. Conclusions

In this paper, we introduced a generative sentence alignment method and a new Arabic–English parallel dataset focusing on legal and literary texts. Our findings reveal the limitations of previous methods. In contrast, the LLM-based methods demonstrated better robustness, maintaining high precision and recall where other methods degraded. Beyond these empirical results, the proposed generative alignment approach provides a flexible and efficient alternative to conventional techniques, while requiring relatively modest computational resources. Consequently, our approach has the potential to support improvements in downstream machine translation (MT) systems, as well as broader applications in cross-lingual transfer and multilingual understanding. Furthermore, the released Arabic–English dataset contributes a domain-specific resource that can support future research in legal and literary translation. Future work can extend this approach to additional low-resource language pairs, and explore multi-agent or self-refinement strategies to further enhance alignment quality.

Limitations

The proposed LLM-based document alignment methods can effectively and accurately align the low-resource Arabic text to English text. However, the required time for the alignment depends on the processing time of the LLMs. In addition, the format of the responses is also essential for the successful parsing and construction of the alignment mappings. Furthermore, the length of the document is also constrained by the context window of the large language models. In our experiments, one simple prompt is used, as described in 10.1 and the effects of advanced prompt methods, such as chain-of-reason (COT), self-reflection, or agentic design, are not explored. Prior work suggests that these methods, especially those involving multi-agent collaboration, such as reflection or reviewer agent mechanisms(Bo et al., 2024) or ensemble mechanisms(Qian et al., 2026), can enhance the performance and reasoning reliability of LLMs. Investigating whether such approaches further improve alignment quality remains an important direction for future research.

In addition, it is well-known that LLMs have difficulties in handling long texts. Future research may explore how long texts, such as an entire novel with hundreds of pages and its translation, can be effectively aligned by LLMs.

7. Acknowledgments

The authors extend their appreciation to Umm Al-Qura University, Saudi Arabia for funding this research work through grant number: 26UQU4350258GSSR01.

8. Funding Statement

This research work was funded by Umm Al-Qura University, Saudi Arabia under grant number: 26UQU4350258GSSR01.

9. Bibliographical References

References

Maha Jarallah Althobaiti. 2022. A simple yet robust algorithm for automatic extraction of parallel sentences: A case study on arabic-english wikipedia articles. *IEEE Access*, 10:401–420.

Xiaohe Bo, Zeyu Zhang, Quanyu Dai, Xueyang Feng, Lei Wang, Rui Li, Xu Chen, and Ji-Rong Wen. 2024. Reflective multi-agent collaboration based on large language models. In *Advances in Neural Information Processing Systems*, volume 37, pages 138595–138631. Curran Associates, Inc.

Peter F. Brown, Jennifer C. Lai, and Robert L. Mercer. 1991. Aligning sentences in parallel corpora. *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics*, pages 169–176.

William A. Gale and Kenneth W. Church. 1993. A program for aligning sentences in bilingual corpora. *Computational Linguistics*, 19(1):75–102.

Philipp Koehn, Vishrav Chaudhary, Ahmed El-Kishky, Naman Goyal, Peng-Jen Chen, and Francisco Guzmán. 2020. Findings of the wmt 2020 shared task on parallel corpus filtering and alignment. In *Proceedings of the Fifth Conference on Machine Translation*, pages 726–742.

Philipp Koehn, Huda Khayrallah, Kenneth Heafield, and Mikel L Forcada. 2018. Findings of the wmt 2018 shared task on parallel corpus filtering. In *EMNLP 2018 Third Conference on Machine Translation (WMT18)*, pages 726–739. Association for Computational Linguistics.

Lei Liu and Min Zhu. 2023. Bertalign: Improved word embedding-based sentence alignment for chinese–english parallel corpora of literary texts. *Digital Scholarship in the Humanities*, 38(2):621–634.

Francesco Maria Molfese, Andrei Stefan Bejgu, Simone Tedeschi, Simone Conia, and Roberto Navigli. 2024. Crocoalign: A cross-lingual, context-aware and fully-neural sentence alignment system for long texts. In *Proceedings of the 18th Conference of the European Chapter of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2209–2220.

Yiyue Qian, Shinan Zhang, Yun Zhou, Haibo Ding, Diego Socolinsky, and Yi Zhang. 2026. Collabeval: Enhancing llm-as-a-judge via multi-agent collaboration.

Rico Sennrich and Martin Volk. 2010. Mt-based sentence alignment for ocr-generated parallel texts. In *Proceedings of the Ninth Conference of the Association for Machine Translation in the Americas (AMTA 2010)*, pages 1–10, Denver, USA.

Wei Shang and Xin Li. 2024. Construction and alignment features of english corpus based on particle swarm optimization algorithm. In *2024 International Conference on Distributed Systems, Computer Networks and Cybersecurity (ICDSCNC)*, pages 1–5. IEEE.

Steve Sloto, Brian Thompson, Huda Khayrallah, Tobias Domhan, Thamme Gowda, and Philipp Koehn. 2023. Findings of the wmt 2023 shared task on parallel data curation. In *Proceedings of the Eighth Conference on Machine Translation*, pages 95–102.

Steinþór Steingrímsson, Hrafn Loftsson, and Andy Way. 2023. SentAlign: Accurate and Scalable Sentence Alignment. In *Proceedings of the 2023 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, Singapore, Singapore. Association for Computational Linguistics.

Brian Thompson and Philipp Koehn. 2019. Vecalign: Improved sentence alignment in linear time and space. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing (EMNLP-IJCNLP)*, pages 1342–1348, Hong Kong, China. Association for Computational Linguistics.

Xiaotian Wang, Takehito Utsuro, and Masaaki Nagata. 2024. Document alignment based on overlapping fixed-length segments. In *Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 4: Student Research Workshop)*, pages 51–61.

Wu Zhang. 2022. Improve sentence alignment by divide-and-conquer. *arXiv preprint arXiv:2201.06907*.

10. Language Resource References

N/A

Appendix

10.1. Prompt Template and Sample Responses

The following prompt template is used in the alignment process.

You are an expert in translation. Align two documents and output bilingual pairs.

Document A is in Arabic, and Document B is in English. Both documents contain sentences that are similar in meaning but may not be direct translations of each other.

Your task is to find the translations from Document B for lines in Document A and output them as bilingual pairs. Use the source line as one sentence as much as possible. In some cases, more than one source lines are matched to more than one target lines.

Find the translations for all the source lines. Do not merge the source lines unless absolutely necessary.

Produce the complete bilingual alignment in one large JSON array.

When finding the translation, ignore the line number such as §0§, §1§ at the beginning of each sentence and keep the line number in the output.

Document A (Arabic):
{arabic_text}

Document B (English):
{english_text}

Output Format:
[{{"Arabic Sentence": "$line no$ نظام القضاء",
"English Sentence": "$line no$Law of the Judiciary" }}]

The sample response is described as follows:

["Arabic Sentence":"§0§ الشركة كيان قانوني يؤسس وفقًا لأحكام النظام بناءً على عقد تأسيس أو نظام أساس يلتزم بمقتضاه شخصان أو أكثر بأن يساهم كل منهم في مشروع يستهدف الربح بتقديم حصة من مال أو عمل أو منهما معًا لاقتسام ما ينشأ عن هذا المشروع من ربح أو خسارة، واستثناء من ذلك، يجوز -وفقًا لأحكام النظام- أن تؤسس الشركة بالإرادة المنفردة لشخص واحد، ويجوز تأسيس شركات غير ربحية وفقًا لما ورد في الباب (السابع) من النظام.",
"English Sentence":"§0§ A company is a legal entity incorporated in accordance with the provisions of this Law pursuant to articles of incorporation or articles of association under which two or more persons undertake to participate in a for-profit enterprise by contributing property or work, or both, to share any profit realized or loss incurred from such enterprise. §1§ As an exception, a company may, under this Law, be incorporated by a single person, and a non-profit company may be incorporated pursuant to the provisions of Part 7 of this Law.",
...
]