Unsupervised Sentence Simplification via Dependency Parsing

Anonymous ACL submission

Abstract

Text simplification is the task of rewriting a text so that it is readable and easily understood. In this paper, we propose a simple yet novel unsupervised sentence simplification system that harnesses parsing structures together with sentence embeddings to produce linguistically effective simplifications. This means our model is capable of introducing substantial modifications to simplify a sentence while maintaining its original semantics and adequate fluency. We establish the unsupervised state-of-the-art at 39.13 SARI on TurkCorpus set and perform competitively against supervised baselines on various quality metrics. Furthermore, we demonstrate our framework's extensibility to other languages via a proof-of-concept on Vietnamese data. Code for reproduction is anonymously published at `https://anonymous.4open.science/r/USDP-744B`.

1 Introduction

Text simplification (TS) contributes to promoting social inclusion by making information more accessible to people with reading comprehension problems such as second-language learners, non-experts or those suffering from cognitive impairment (Siddharthan, 2014; Stajner, 2021). Simplification can take many forms. It may involve **splitting** and/or **reordering** parts of the text to produce simpler syntactic structures. **Lexical transformation** can also be performed through the substitution of easier and more familiar vocabularies. While **deleting** redundant details produces an output shorter in length, the output can also become longer with extra information added to provide explanations for difficult concepts, i.e., **elaboration**. Regardless of the operations, the simple variants must preserve the key meaning of the original texts.

Previous studies criticize existing systems for being opaque, suboptimal and semantically compromising (Garbacea et al., 2021; Maddela et al., 2021; Stajner, 2021). Table 1 illustrates how the meaning is easily sacrificed with respect to minor changes to the sentence structure. Generally, TS works must be mindful of the trade-off of **Simplicity** vs. **Fluency** and **Adequacy** (semantics similarity). Although some works explicitly incorporate all three properties in the training objectives (Laban et al., 2021; Zhang and Lapata, 2017; Kumar et al., 2020), their evaluations do not clearly explain whether or not simplicity is induced as the cost of fluency and adequacy. By exploiting deep dependency parsing, we contribute a novel unsupervised strategy called **Family Sampling** that strictly enforces grammatically fluent simplification while ensuring the most important ideas are retained. We shed light on how our model achieves the balance of these three properties through both automatic metrics and human judgement, which at the same time substantiates our superiority over unsupervised counterparts.

Original sentence	Simple variant
In Ethiopia, HIV disclosure is low	In Ethiopia , HIV is low
Mustafa Shahbaz , 26 , was shopping for books about science	Mustafa Shahbaz, 26 years old , was a group of books about science
Healthy diet linked to lower risk of chronic lung disease	Healthy diet linked to lung disease

Table 1: Examples of logic errors produced by `ACCESS` (Martin et al., 2020) and `DMLMTL` (Guo et al., 2018), taken from (Garbacea et al., 2021).

Whereas most models are restricted to the language of the data they are trained on, we demonstrate that our framework readily extends to other languages by adapting it to simplify Vietnamese texts. We also address interpretability by providing linguistically-motivated empirical evidence confirming the intuition behind our framework.

2 Related Work

Supervised TS. Earlier works inherit techniques from statistical machine translation (Brown et al., 1990) to translate a text of the *complex* language to the *simple* language. The translation model is learned through aligned words or phrases in normal-simplified text pairs, referred to as phrase-based simplification (Coster and Kauchak, 2011; Koehn et al., 2007; Narayan and Gardent, 2014; Wubben et al., 2012). Alternatively, in syntax-based simplification (Zhu et al., 2010), the alignment units are syntactic components. The first neural sequence-to-sequence text simplification system is proposed in (Nisioi et al., 2017). Utilizing the same architecture, other works (Guo et al., 2018; Zhang and Lapata, 2017) further employ reinforcement learning with a reward function that is a weighted sum of three component rewards: simplicity, relevance and fluency. Audience Centric Sentence Simplification `ACCESS` is a recent supervised state-of-the-art approach (Martin et al., 2020) that conditions the simplified outputs on different attributes of text complexity. These models rely on parallel corpora to implicitly learn hybrid transformation patterns. Despite impressive results, the scarcity of high-quality and large-scale datasets heavily impedes progress in supervised TS (Alva-Manchego et al., 2020). The attention has thus been shifted towards semi-supervised and unsupervised approaches.

Semi-Supervised TS. Instead of using aligned data, Zhao et al. (2020) introduce a noising mechanism to generate parallel examples from any English dataset, then train denoising autoencoders to reconstruct the original sentences. Under a framework called **back-translation**, Martin et al. (2021) use multilingual translation systems to produce various simpler paraphrases from monolingual corpora (e.g., English to French, then French to English), thus eliminating the need for labeled data. In the same spirit, Mallinson et al. (2020) recently adopt a cross-lingual strategy to simplify low-resource languages (e.g., German) from high-resources languages (e.g., English). Instead of having translation done in the data pre-processing step, they develop a system that performs simplification and translation at the same time.

Unsupervised TS. The first unsupervised neural model for text simplification is proposed by Surya et al. (2019). The model is trained to minimize adversarial losses on two separate sets of complex and simple sentences extracted from a parallel Wikipedia corpus. `DisSim` (Niklaus et al., 2019) is another effort focusing on splitting and deletion by applying 35 hand-crafted grammar rules over a constituency parse tree. Recent works tend to favor edit and decoding-based approaches. This line of work is advantageous since not only can the system generate hybrid outputs without relying on aligned datasets, but it can also allow for quality control explicitly via a scoring function balancing simplicity, fluency and semantics preservation. The algorithm of Kumar et al. (2020) iteratively edits a given complex sentence to make it simpler using four operations: removal, extraction, reordering and substitution, while (Kariuk and Karamshuk, 2020) implement beam search with simplicity-aware penalties for sentence simplification. In the same setting as (Zhang and Lapata, 2017), `KiS` (Laban et al., 2021) revisits reinforcement learning and tackles simplification for paragraphs without supervision. However, the method involves end-to-end training on multiple Transformer-based models, which is computationally expensive and challenging to extend to new settings.

Contributions. In contrast to the above, we propose a light-weight solution for sentence TS making full use of pre-training, i.e., neither fine-tuning nor end-to-end training is required, thereby making it highly reproducible and robust to out-of-distribution examples. Specifically, we improve on the existing decoding procedure through a linguistics-based unsupervised framework. We perform structural and lexical simplification sequentially, rather than simultaneously like previous works since it would support interpretation and allow for controllability. This sequential approach is also adopted in (Maddela et al., 2021), which leverages `DisSim` together with a self-designed paraphrasing system. We first develop a **stand-alone decoding framework** for structural simplification, then adopt **back-translation** for lexical simplification and paraphrasing. After studying prior works, we find that **splitting** and **elaboration** are difficult to implement without labeled data or heavily injected grammar rules, whereas our goal is to maximize the capacity of TS system in the absence of external knowledge. An interesting discovery is that back-translation in phase 2 is a convenient technique, in that it can also perform **reordering** (as part of the rewriting process), if the operation is necessary to produce a familiar structure. Thus, in the first phase, we choose to

focus on **deletion** - the operation that easily leads to poor adequacy if not carefully done. Although we only tackle sentence simplification, in **Section 7**, we provide directions on how our framework can be flexibly adapted for various purposes, including paragraph-level simplification.

3 Method

We propose a two-phase pipeline that tackles syntactic and lexical simplification one by one. The first phase consists of an independent left-to-right decoder operating in a much more efficient search space induced by dependency relations among words in a sentence. Though this phase is mainly about deletion, during the process, the system can also perform **chunking**, i.e., breaking a sentence into meaningful phrases. In the second phase, we back-translate the generated English outputs from phase 1 to generate effective paraphrases and lexical simplifications.

3.1 Structural Simplification

3.1.1 Search Objective

Given an input sentence $c := (c_1, c_2, ..., c_n)$, we aim to generate a shorter sentence $s := (s_1, s_2, ..., s_m)$ expressing the same meaning as c. Whereas the previous works perform deletions by imposing length constraints, we go beyond length reduction and strictly define which parts of a sentence to keep and which to remove. The goal is to eliminate redundant details – those if removed do not significantly alter the meaning of the entire sentence. We quantify the importance of a word by measuring local changes in semantic similarity scores when omitting it from the original sentence. This motivates our search objective function as follows

$$f(s) = f_{sim}(c, s) + \alpha f_{flu}(s) + f_{depth}(s) \quad (1)$$

where α is the relative weight on Fluency score. The reason why the weights of score functions f_{sim} and f_{depth} are the same is explained under **Section 4.2**. The decoding objective is a linear combination of individual scoring functions with each score normalized within the range $[0, 1]$. Details of each function are described below.

Semantic Similarity $f_{sim}(c, s)$. Using cosine distance as a proxy for semantic relevance has been widely adopted across TS works. We calculate cosine similarity between sentence embeddings of c and generated hypothesis $s_{1:t}$ at each time step t. We utilize the pre-trained sentence-BERT model (SBERT) (Reimers and Gurevych, 2020a), which is best known for its superiority in producing semantically meaningful sentence embeddings, whereas other unsupervised works use weighted average of word embeddings (Kumar et al., 2020; Schumann et al., 2020; Zhao et al., 2020) or LSTM-encoded hidden representations (Zhang and Lapata, 2017).

$$f_{sim}(c, s_{1:t}) = cos(e_c, e_{s_{1:t}})$$

The intuition is, if a candidate word i is more important than another word j, add i to the sentence will increase the similarity score more than when adding j. It is observed that SBERT sentence embeddings capture this behavior, and exactly how it works is explained in **Section 6**.

Fluency $f_{flu}(s)$. Our fluency scorer quantifies the grammatical accuracy of a sequence based on a constituent-based 4-gram language model, given as

$$f_{flu}(s_{1:t}) = \frac{1}{|s_{1:t}|} \sum_{u=1}^{t} \log p(pos_u \,|\, pos_{1:u-1})$$

where pos_t indicates the part-of-speech of token s_t The language model is pre-trained on a massive unlabeled corpus. Because English constituents are bounded, constituent-based language model is a reusable light-weight solution compared to regular vocabulary-based language models.

Tree Depth Constraint $f_{depth}(s)$. Dependency tree depth is a popular measure of syntactic complexity in various literature in linguistics (Genzel and Charniak, 2003; Sampson, 1997; Xu and Reitter, 2016). A deeper tree contains more dependencies indicating complex structures e.g., usage of subordinate clauses. ACCESS (Martin et al., 2020) has recently provided empirical evidence showing that controlling maximum depth of dependency tree yields the most effective syntactical simplifications. Thus, f_{depth} further scores candidate sentences by the **inverse maximum tree depth** reached at the generated token. This constraint prevents the decoder from going too deep, thereby producing a structurally simpler output.

3.1.2 Search Space

Deletion is a form of extractive summarization by nature, motivating us to adopt the word-extraction method proposed by (Schumann et al., 2020). They suggest candidates be selected from tokens in the input sentence, instead of the corpus vocabulary.

Specifically at each step, a new candidate is sampled from words in the input sentence that are not in the current summary. We argue that this is not necessary and propose a more efficient approach. Each token exists in a directed relation with its parent (or head), which determines the grammatical function of that token. The head-dependent relation is also an approximation for the semantic relationship between them (Jurafsky and Martin, 2014). At each step, we thus only consider tokens that are the children of the previously generated token, resulting in a smaller search space. We additionally arrange the words in the same order as the input before scoring hypotheses since word order plays a critical role both grammatically and semantically. This approach allows us to achieve optimal solution while using a small beam size. We refer to this novel strategy as **Family Sampling**.

3.1.3 Search Algorithm

Figure 1 provides a running example for our algorithm. We integrate our novel family sampling strategy with a regular beam search algorithm that keeps top k hypotheses with highest scores. To begin with, we condition the sequence on the main subject of the sentence, i.e. the subject of the ROOT verb. This does not affect the rest of the sentence, but contributes to simplification by directly introducing the main verb and subject. For each search step, a candidate token s_t is sampled from the family of child nodes of token s_{t-1}, excluding those having been previously generated. We score each candidate according to (1) and select k hypotheses with the highest scores for the next generation step. Our search terminates when the output sentence reaches a predefined length λ and satisfies the minimum similarity threshold τ. We wish to find the shortest most similar sub-sequence, and the intuition is to preserve as much semantics as possible by keeping tokens that add the most semantic value to the sequence. As mentioned in (Schumann et al., 2020), given input length n, target output length m and corpus vocabulary size V, auto-regressive or edit-based generation has search space of V^m, while ours is restricted to C_n^m. Regarding time complexity, our algorithm is bounded by $O(d \times k \times max_ch(s_{t-1}))$ with parsing tree depth d, beam size k and $max_ch(s_{t-1}))$ being the maximum number of children of a token s_{t-1} where $ch(s_{t-1}) \leq n$ and

Chunking. A branch in the tree corresponds to a meaningful chunk in the sentence. A <SEP> token is added to induce a chunk whenever a sampling set is empty. Humans naturally perform simplification in this manner by chunking a complex sentence into understandably simpler structures. Our results show that most of these chunks tend to be prepositional and adjective phrases. After a <SEP>, we reverse the tree and restart family sampling with the token nearest to ROOT.

3.2 Back Translation

Phase 2 consists of two reliable off-the-shelf machine translation systems English-X and X-English where X can be any other language. We simply translate the structurally simpler sentences in English to X, then have the outputs back-translated to English. This technique is applied in style transfer tasks (Prabhumoye et al., 2018; Zhang et al., 2018) to disentangle the content and stylistic characteristics of the text. Thus, we rely on back-translation to strip the complex style off a text while keep meaning unchanged. Not only does it enhance the quality of our simplifications via paraphrasing and lexical substitution, but we also find it particularly useful to correct subpotimality in the decoder's output. Another advantage of this technique is that one can further collect various paraphrases by exploiting multiple languages.

4 Experiments

4.1 Data

We evaluate our English model on TurkCorpus (Xu et al., 2016) and PWKP (Zhang and Lapata, 2017). TurkCorpus is a standard dataset for evaluating sentence simplification works, originally extracted from WikiLarge corpus compiled from (Zhang and Lapata, 2017). It contains 2000 sentences for validation and 359 for testing, each has 8 simplification references collected through crowd-sourcing. PWKP is the test set of WikiSmall - another dataset constructed from main-simple Wikipedia articles. PWKP provides 100 test sentences with 1-to-1 aligned reference. Newsela (Xu et al., 2015) is another commonly used dataset in text simplification works, which is unfortunately unavailable due to restricted access rights. Without training, we directly run our model and evaluate the outputs on TurkCorpus and PWKP test sets, and to prove the robustness of our system, we further use CNN/Daily Mail, a different dataset (See et al., 2017) for training the fluency model, leaving both evaluation corpora untouched. For the Vietnamese model, we train the

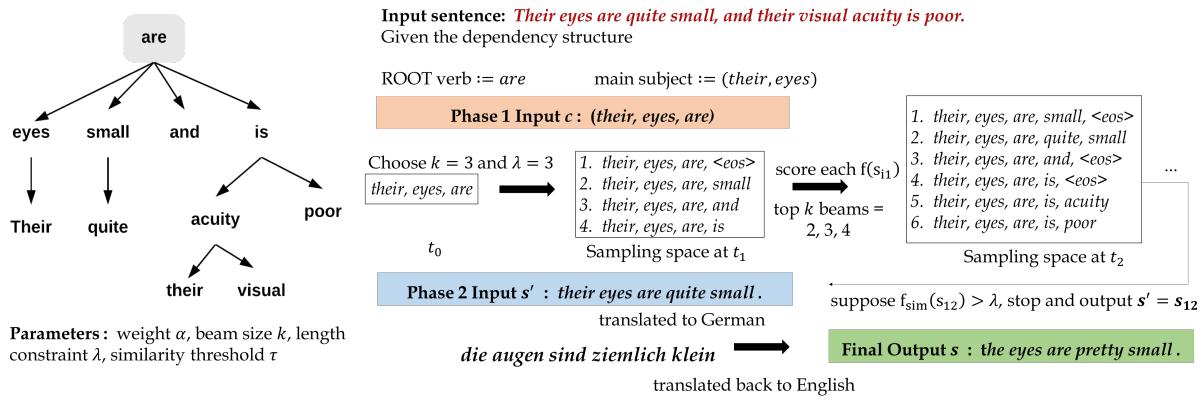


Figure 1: At each search step, a candidate token is sampled from the direct children of previously generated token. Score each candidate according to the objective function and select k hypotheses with highest scores for the next generation step. The decoder terminates once there is a sequence reaching length λ with at least similarity τ.

fluency model on the public Vietnamese news corpus CP_Vietnamese-UNC of 41947 sentences, then generate simplifications of 200 sentences extracted from Vietnamese law corpus CP_Vietnamese-VLC in an unsupervised manner. Both datasets are open sourced by Underthesea NLP [1].

4.2 System Details

We utilize SpaCy[2] and Berkeley Neural Parser (Kitaev and Klein, 2018) for constituent and dependency parsing. Since SpaCy does not support Vietnamese, we parse Vietnamese texts through VnCoreNLP[3] (Vu et al., 2018). VnCoreNLP is built upon Vietnamese Treebank (Nguyen et al., 2009), which contains 10,200 constituent trees formatted similarly to Penn Treebank (Marcus et al., 1993). We obtain sentence embeddings from SBERT model `paraphrase-mpnet-base-v2` (Reimers and Gurevych, 2020b) for English, and the multilingual version `distiluse-base-multilingual-cased-v2` for Vietnamese (Reimers and Gurevych, 2020a). Our fluency model is a 4-gram language model with Kneser-Ney smoothing implemented via NLTK[4] package. Before that, we randomly sample 1 million sentences from CNN/Daily Mail set and parse them into sequences of constituents. For example, the sentence *Their eyes are small* is transformed into `PRON NOUN AUX ADJ`.

The back-translation procedure employs free Google Translate service[5] - a robust neural translation system that support two-way translation **English-German** and **German-English**. Though the system can be run with any available target languages, German is selected for our English model because it is a well-resourced language. For the Vietnamese model, we simply use English as the target language. While our search algorithm can automatically guarantee language idiomaticity, grammatical accuracy is an issue since the model tends to prefer content words to function words to maximize semantic similarity. Thus, we strictly force Fluency to be twice as important, i.e., $\alpha = 2$, while set equal weights to Similarity and Depth constraint. Our base model is evaluated at $\lambda = 0.5$ and $\tau = 0.95$, reported as `USDP-Base`. In order to validate our effectiveness more comparably, we produce additional variants approximating τ to same level of competing unsupervised methods, respectively at $\tau = 0.90$ for TurkCorpus (`USDP-Match`^{a}) and $\tau = 0.75$ for PWKP (`USDP-Match`^{b}). Across all experiments, **beam size is fixed at 5**, and in order to understand the effect of back-translation, we also evaluate the quality of simplifications before and after phase 2.

4.3 Competing Models

We benchmark our system against existing supervised and unsupervised English sentence simplification models. Supervised systems include `PBMT-R` (Wubben et al., 2012), `SBMT-SARI` (Xu et al., 2016), `Dress` / `Dress-Ls` (Zhang and Lapata, 2017) and recent state-of-the-art `ACCESS` (Martin et al., 2020). We also consider semi-supervised `BTTS` / `BTRLTS` / `BTTS100` (Zhao et al., 2020) and unsuper-

[1] `github.com/undertheseanlp/resources`
[2] `spacy.io`
[3] `github.com/vncorenlp/VnCoreNLP`
[4] `nltk.org`
[5] `translate.google.com`

vised counterparts UNTS / UNTS10K (Surya et al., 2019) and RM+EX / RM+EX+LS / RM+EX+RO / RM+EX+LS+RO (Kumar et al., 2020).

5 Results

5.1 Automatic Evaluation

We use the EASSE package (Alva-Manchego et al., 2019; Martin et al., 2018) to compute standardized simplification metrics and perform evaluation on publicly accessible outputs of competing systems. These include Compression ratio (**CR**), Exact copies (**CP**), Split ratio (**%SP**), Additions proportion (**%A**) and Deletions proportion (**%D**), all of which are evaluated against the source sentences. Details on these automatic metrics can be found in Appendix A. We exclude BLUE and FKGL since BLEU is previously reported to be a poor estimate of simplicity and FKGL only applies to text of at least 200 words (Alva-Manchego et al., 2020; Wubben et al., 2012; Xu et al., 2016). We add to the current simplification evaluation suite the measures of semantic similarity and fluency. Similarity score is again based on cosine similarity between sentence embedding vectors (**SIM**), and we adopt a "third-party" language model for scoring fluency (**FL**). This is to assure fair comparison among systems since ours has a different fluency scoring scheme. We use **pseudo-log-likelihood scores** (PPLs) proposed in (Salazar et al., 2020), which is shown to promote linguistic fluency rather than pure likelihood in conventional log probabilities. We also evaluate the reference sentences on these quality metrics, and benchmark the system outputs against them through average **SARI** and component **Add**, **Keep**, **Del** scores (Xu et al., 2016). This however is done only on TurkCorpus set since PWKP only provides 1 reference. Table 2, 3 and 4 present results of automatic evaluation respectively on TurkCorpus, PWKP and CP_Vietnamese-VLC, both before and after back-translation (**BT**).

5.1.1 TurkCorpus

We establish the unsupervised state-of-the-art SARI on TurkCorpus with **+1.65** point improvement over the closest baseline and only behind two supervised methods: ACCESS and SBMT-SARI. In addition to the competitive performance on Compression ratio and Split ratio, we outperform the current semi-supervised and unsupervised across all other quality metrics. Our simplifications have the highest fluency at **-2.47** and similarity score at **0.95** while achieving remarkably high percentages of additions at **16%** and deletions of **21-25%** at the same level of some supervised methods. Our raw outputs from phase 1 alone gains fairly high proportions of deletions as lowering τ. For other methods, this number generally takes both deletions and substitutions into account, but in this phase, it reflects our model's effectiveness in performing deletions since substitutions are not implemented until phase 2. Figures 4 and 5 in Appendix C visualize how well our base model balances simplicity with adequacy and fluency compared to other methods. Samples of system outputs are additionally provided in Appendix D.

5.1.2 PWKP

As Kumar et al. (2020) do not experiment on PWKP, we run their codes to evaluate RM+EX+LS+RO model on PWKP set for comparison. Overall, our model and RM+EX+LS+RO produce more diverse simplifications than the supervised systems, measured by significantly higher proportion of additions and deletions. Interestingly, the quality of unsupervised outputs is also closer to that of reference outputs, in which RM+EX+LS+RO achieves consistently better performance. This is probably because the model is accompanied by a pre-trained Word2Vec on WikiLarge data, which has a relatively similar distribution to PWKP as both are Wikipedia-based. Meanwhile, none of our variants see any similar examples of any kind beforehand. Given such a high level of modification, we again have the highest similarity score at **0.96**, and when we try to match the similarity level as in USDP-Match^b, we achieve more compression at **54%** and deletions at **56%** while preserving slightly higher semantics than RM+EX+LS+RO. The fluency scores of outputs from all automated systems remain far behind human simplifications.

5.1.3 CP_Vietnamese-VLC

As a proof-of-concept for our approach in another language, we only conduct evaluation on USDP-Base. We do not report PPLs since that model has only been shown to work on English data. We instead report **LevSIM**, normalized character-level Levenshtein similarity (Levenshtein, 1966), which demonstrates that the output structures do not deviate significantly from the original. Results in Table 4 are consistent with what we have achieved on English corpora, proving the po-

TurkCorpus	CR↓	CP↓	%SP↑	%A↑	%D↑	FL↑	SIM↑	SARI↑	Add↑	Keep↑	Del↑
Reference	0.95	1.07	0.16	0.14	0.18	-2.63	**0.95**	-	-	-	-
Supervised											
Dress	**0.75**	0.22	0.99	0.04	**0.27**	-2.66	0.91	36.84	2.5	65.65	42.36
Dress-Ls	**0.77**	0.26	0.99	0.04	**0.26**	-2.63	0.92	36.97	2.35	67.23	41.33
PBMT-R	0.95	0.11	1.03	0.10	0.11	-2.59	**0.96**	**38.04**	5.04	**73.77**	35.32
ACCESS	0.94	**0.04**	**1.20**	**0.16**	0.16	-2.52	**0.95**	**41.38**	**6.58**	72.79	44.78
SBMT-SARI	0.94	0.10	1.02	**0.16**	0.13	-2.65	**0.96**	**39.56**	5.46	72.44	40.76
Semi-Supervised											
BTTS100	0.92	0.45	1.02	0.03	0.10	**-2.46**	**0.97**	34.48	1.51	**74.44**	27.48
BTTS	0.92	0.20	1.17	0.08	0.14	-2.66	**0.96**	36.38	1.9	71.03	36.22
BTRLTS	0.92	0.19	1.16	0.08	0.15	-2.70	**0.96**	36.49	2.14	70.31	37.03
Unsupervised											
UNTS	0.85	0.21	1.00	0.06	0.17	-2.70	0.89	36.29	0.83	69.44	38.61
UNTS_10K	0.88	0.19	1.01	0.07	0.14	-3.10	0.92	37.15	1.12	71.34	38.99
RM+EX	0.83	0.44	1.00	0.01	0.15	-2.58	0.94	35.88	0.84	73.14	33.65
RM+EX+LS	0.82	0.16	1.00	0.06	**0.21**	-2.91	0.90	37.48	1.59	68.20	42.65
RM+EX+RO	0.86	0.36	1.01	0.02	0.14	-2.61	0.94	36.07	0.99	72.36	34.86
RM+EX+LS+RO	0.85	0.13	1.01	0.08	**0.20**	-2.92	0.90	37.27	1.68	67.00	43.12
Our system											
USDP-Base											
With BT	0.92	**0.04**	1.01	**0.16**	**0.21**	**-2.47**	**0.95**	**39.13**	**6.77**	64.44	**46.19**
Without BT	0.95	0.15	1.00	0.07	0.09	-2.88	**0.98**	34.13	0.87	71.34	30.18
USDP-Match[a]											
With BT	0.88	**0.04**	1.01	**0.15**	**0.25**	-2.55	0.94	**38.33**	**6.28**	62.13	**46.58**
Without BT	0.89	0.13	1.00	0.07	0.15	-3.05	**0.96**	34.44	0.94	68.57	33.82

Table 2: Results on TurkCorpus. ↑ Higher is better. ↓ Lower is better.

PWKP	CR↓	CP↓	%SP↑	%A↑	%D↑	FL↑	SIM↑
Reference	0.81	0.03	**1.31**	**0.17**	0.32	**-1.39**	0.91
Supervised							
Dress	0.62	0.11	1.01	0.02	0.39	-2.18	0.87
Dress-Ls	0.63	0.13	1.01	0.01	0.37	-2.10	0.88
PBMT-R	0.96	0.14	1.01	0.06	0.07	-2.05	**0.97**
Unsupervised							
RM+EX+LS+RO	0.61	**0.01**	**1.21**	**0.17**	**0.52**	-2.68	0.81
Our system							
USDP-Base							
With BT	0.87	0.03	1.00	**0.16**	0.28	-2.05	**0.96**
Without BT	0.88	0.08	1.00	0.06	0.15	-2.64	**0.95**
USDP-Match[b]							
With BT	**0.54**	**0.00**	1.00	0.11	**0.56**	-2.38	0.84
Without BT	**0.53**	0.03	1.00	0.03	0.49	-3.36	0.85

Table 3: Results on PWKP. ↑ Higher is better. ↓ Lower is better.

tential to apply the framework to other languages.

5.2 Human Evaluation

Human judgement is critical to assess text generation. We randomly select 50 sentences from TurkCorpus test set, and have 5 volunteers (2 native and 3 non-native speakers with adequate English proficiency) examine the simplified outputs from ACCESS (supervised state-of-the-art), RM+EX+LS (closest and best performing unsupervised variant) and our method USDP-Base. In a similar setup to the previous studies (Kumar et al., 2020; Maddela et al., 2021; Zhao et al., 2020), the volunteers are asked to provide ratings for each simplification version on 3 dimensions: **Simplicity**, **Adequacy** and **Fluency**. We also have 50 Vietnamese simplifications from CP_Vietnamese-VLC outputs assessed by 4 native Vietnamese speakers on the same quality dimensions. We simply report the average ratings in Table 5, substantiating that we surpass both ACCESS and RM+EX+LS on all dimensions, and our simplified sentences in Vietnamese are perceived to have adequate quality.

5.3 Controllability

Table 6 displays output results on different values of τ and λ. Adjusting similarity threshold τ has more impact on the output quality than length ratio λ. This is simply because our algorithm must satisfy a pre-defined τ before termination, regardless of length constraints. This shows that lowering similarity threshold encourages the model to produce more deletions and compression, which however does not occur at the cost of semantics preservation. Even when τ is set to 0.70, the output sentences still have very high similarity scores. Little content is lost since we not only reduce length but also seek to maximize semantics preservation by extracting important tokens only, most of which turn out to be content words. This behavior is discussed in detail in the next section.

CP_Vietnamese-VLC	CR↓	CP↓	%SP↑	%A↑	%D↑	LevSIM↑	SIM↑
With BT	0.89	0.00	1.06	0.11	0.20	0.86	0.91
Without BT	0.91	0.00	0.99	0.06	0.12	0.91	0.94

Table 4: Results of `USDP-Base` on CP_Vietnamese-VLC

Model	Fluent	Adequate	Simple
English			
`USDP-Base`	**4.32**	**3.93**	**3.22**
`ACCESS`	4.16	3.46	3.18
`RM+EX+LS`	3.59	3.12	2.86
Vietnamese			
`USDP-Base`	3.33	3.48	3.04

Table 5: Human Evaluation Results on TurkCorpus (English) and CP_Vietnamese-VLC (Vietnamese) datasets.

Value	CR↓	%D↑	SARI↑	SIM↑
Effect of λ at $\tau = 0.95$				
0.25	0.95	0.08	33.60	0.98
0.50	0.95	0.08	33.60	0.98
0.75	0.96	0.07	33.47	0.98
1.00	0.99	0.04	32.86	0.98
Effect of τ at $\lambda = 0.5$				
0.70	0.81	0.22	33.63	0.92
0.80	0.84	0.19	33.77	0.94
0.90	0.91	0.12	33.80	0.97
0.95	0.95	0.08	33.60	0.98

Table 6: Effects of threshold values on simplification quality of 100 sentences from TurkCorpus. Both are evaluated on `USDP-Base`.

6 Discussion

In the first phase, we ensure each added token brings about significant improvement in semantics. A syntactic investigation was conducted to understand this behavior better, illustrated in Figure 6 in Appendix C. Figure 6*a* first examines the correlation between the tokens' part-of-speech and changes in similarity on average. After randomly sampling 1 million English sentences from all the data, for each sentence, we consecutively remove every token and track how much the SBERT embedding similarity score is reduced, then add the token back in before evaluating a new one. **Content words** including `NOUN`, `PRON`, `VERB`, `ADJ` and `ADV`, which are deemed semantically more important, each causes more than 6% reduction in semantic value when omitted. Meanwhile, **function words** such as `CONJ` or `DET` only reduces similarity scores by less than 4% each. This implies that a child token serving as an adposition, for instance, is less likely to be selected in the sampling set compared to when it is a noun or adjective, especially in the family with more members than the chosen beam size. Hence, using the family sampling algorithm, the decoder is likely to eliminate the entire branch corresponding to prepositional or adverbial phrases in the dependency tree, as reported in Appendix D. We also examine the effect of tree depth and sentence length (word count) on similarity changes, which is discussed in (Schumann et al., 2020) as a problem of position bias. We find that this is not a major issue to our work. Although a large portion of important content is allocated towards the top of the parsing tree as well as the beginning of the sentence (Figure 6*b*), the distributions of content words and function words are almost uniform across the tree (Figure 6*c*). We therefore rule out position bias and instead attribute this to human nature of writing.

With respect to the second phase, the role of back-translation is to introduce meaningful diversity, and we observe that back-translation does more paraphrasing than simple lexical substitution. Therefore, sometimes the output sentences must be longer to be re-written in a simpler way, resulting in slightly less compression (i.e., relatively high Compression ratio).

7 Conclusion and Future Work

We implement the novel **family sampling** strategy on top of the regular beam-search-based decoding for sentence simplification. Our work directly tackles the data scarcity issue by proposing an unsupervised framework that generates hybrid outputs in a light-weight architecture.

Our mechanism can be adapted to paragraph-level simplification by measuring semantic changes to the paragraph given the removal of any sentence. In this case, the importance of a sentence should also factor in the co-referent relation with other sentences as proposed in (Liu et al., 2021). Since we already have phrasal chunks, we can additionally investigate where reordering or deleting any of them would result in a simpler tree structure without significant semantic reduction. Furthermore, our proof-of-concept of Vietnamese simplification demonstrates it has scope for improvement, and that the framework can also be applied to other languages with similar dependency structures.

References

Fernando Alva-Manchego, Louis Martin, Antoine Bordes, Carolina Scarton, Benoît Sagot, and Lucia Specia. 2020. ASSET: A dataset for tuning and evaluation of sentence simplification models with multiple rewriting transformations. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 4668–4679, Online. Association for Computational Linguistics.

Fernando Alva-Manchego, Louis Martin, Carolina Scarton, and Lucia Specia. 2019. EASSE: Easier automatic sentence simplification evaluation. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP): System Demonstrations*, pages 49–54, Hong Kong, China. Association for Computational Linguistics.

Peter F. Brown, John Cocke, Stephen A. Della Pietra, Vincent J. Della Pietra, Fredrick Jelinek, John D. Lafferty, Robert L. Mercer, and Paul S. Roossin. 1990. A statistical approach to machine translation. *Computational Linguistics*, 16(2):79–85.

William Coster and David Kauchak. 2011. Simple English Wikipedia: A new text simplification task. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*, pages 665–669, Portland, Oregon, USA. Association for Computational Linguistics.

Cristina Garbacea, Mengtian Guo, Samuel Carton, and Qiaozhu Mei. 2021. Explainable prediction of text complexity: The missing preliminaries for text simplification. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1086–1097, Online. Association for Computational Linguistics.

Dmitriy Genzel and Eugene Charniak. 2003. Variation of entropy and parse trees of sentences as a function of the sentence number. In *Proceedings of the 2003 Conference on Empirical Methods in Natural Language Processing*, pages 65–72.

Han Guo, Ramakanth Pasunuru, and Mohit Bansal. 2018. Dynamic multi-level multi-task learning for sentence simplification. In *Proceedings of the 27th International Conference on Computational Linguistics (COLING 2018)*.

Dan Jurafsky and James H Martin. 2014. Speech and language processing. *US: Prentice Hall*.

Oleg Kariuk and Dima Karamshuk. 2020. Cut: Controllable unsupervised text simplification. *arXiv e-prints*, pages arXiv–2012.

Nikita Kitaev and Dan Klein. 2018. Constituency parsing with a self-attentive encoder. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 2676–2686, Melbourne, Australia. Association for Computational Linguistics.

Philipp Koehn, Hieu Hoang, Alexandra Birch, Chris Callison-Burch, Marcello Federico, Nicola Bertoldi, Brooke Cowan, Wade Shen, Christine Moran, Richard Zens, Chris Dyer, Ondřej Bojar, Alexandra Constantin, and Evan Herbst. 2007. Moses: Open source toolkit for statistical machine translation. In *Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics Companion Volume Proceedings of the Demo and Poster Sessions*, pages 177–180, Prague, Czech Republic. Association for Computational Linguistics.

Dhruv Kumar, Lili Mou, Lukasz Golab, and Olga Vechtomova. 2020. Iterative edit-based unsupervised sentence simplification. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7918–7928. Association for Computational Linguistics.

Philippe Laban, Tobias Schnabel, Paul N. Bennett, and Marti A. Hearst. 2021. Keep it simple: Unsupervised simplification of multi-paragraph text. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics*, volume 1.

Vladimir Iosifovich Levenshtein. 1966. Binary codes capable of correcting deletions, insertions and reversals. *Soviet Physics Doklady*, 10(8):707–710. Doklady Akademii Nauk SSSR, V163 No4 845-848 1965.

Jingzhou Liu, Dominic JD Hughes, and Yiming Yang. 2021. Unsupervised extractive text summarization with distance-augmented sentence graphs. In *Proceedings of the 44th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 2313–2317.

Mounica Maddela, Fernando Alva-Manchego, and Wei Xu. 2021. Controllable text simplification with explicit paraphrasing. In *Proceedings of the North American Association for Computational Linguistics (NAACL)*.

Jonathan Mallinson, Rico Sennrich, and Mirella Lapata. 2020. Zero-shot crosslingual sentence simplification. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 5109–5126, Online. Association for Computational Linguistics.

Mitchell P. Marcus, Beatrice Santorini, and Mary Ann Marcinkiewicz. 1993. Building a large annotated corpus of English: The Penn Treebank. *Computational Linguistics*, 19(2):313–330.

Louis Martin, Éric de la Clergerie, Benoît Sagot, and Antoine Bordes. 2020. Controllable sentence simplification. In *Proceedings of the 12th Language Resources*

and Evaluation Conference, pages 4689–4698, Marseille, France. European Language Resources Association.

Louis Martin, Angela Fan, Éric de la Clergerie, Antoine Bordes, and Benoît Sagot. 2021. Muss: Multilingual unsupervised sentence simplification by mining paraphrases. *arXiv preprint arXiv:2005.00352*.

Louis Martin, Samuel Humeau, Pierre-Emmanuel Mazaré, Éric de La Clergerie, Antoine Bordes, and Benoît Sagot. 2018. Reference-less quality estimation of text simplification systems. In *Proceedings of the 1st Workshop on Automatic Text Adaptation (ATA)*, pages 29–38, Tilburg, the Netherlands. Association for Computational Linguistics.

Shashi Narayan and Claire Gardent. 2014. Hybrid simplification using deep semantics and machine translation. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 435–445, Baltimore, Maryland. Association for Computational Linguistics.

Phuong-Thai Nguyen, Xuan-Luong Vu, Thi-Minh-Huyen Nguyen, Van-Hiep Nguyen, and Hong-Phuong Le. 2009. Building a large syntactically-annotated corpus of Vietnamese. In *Proceedings of the Third Linguistic Annotation Workshop (LAW III)*, pages 182–185, Suntec, Singapore. Association for Computational Linguistics.

Christina Niklaus, Matthias Cetto, André Freitas, and Siegfried Handschuh. 2019. Transforming complex sentences into a semantic hierarchy. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 3415–3427, Florence, Italy. Association for Computational Linguistics.

Sergiu Nisioi, Sanja Štajner, Simone Paolo Ponzetto, and Liviu P. Dinu. 2017. Exploring neural text simplification models. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 85–91, Vancouver, Canada. Association for Computational Linguistics.

Shrimai Prabhumoye, Yulia Tsvetkov, Ruslan Salakhutdinov, and Alan W Black. 2018. Style transfer through back-translation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 866–876, Melbourne, Australia. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020a. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4512–4525, Online. Association for Computational Linguistics.

Nils Reimers and Iryna Gurevych. 2020b. Making monolingual sentence embeddings multilingual using knowledge distillation. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics.

Julian Salazar, Davis Liang, Toan Q. Nguyen, and Katrin Kirchhoff. 2020. Masked language model scoring. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 2699–2712, Online. Association for Computational Linguistics.

Geoffrey Sampson. 1997. Depth in English grammar. *Journal of Linguistics*, 33(1):131–151.

Raphael Schumann, Lili Mou, Yao Lu, Olga Vechtomova, and Katja Markert. 2020. Discrete optimization for unsupervised sentence summarization with word-level extraction. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5032–5042.

Abigail See, Peter J. Liu, and Christopher D. Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1073–1083, Vancouver, Canada. Association for Computational Linguistics.

Advaith Siddharthan. 2014. A survey of research on text simplification. *ITL-International Journal of Applied Linguistics*, 165(2):259–298.

Sanja Stajner. 2021. Automatic text simplification for social good: Progress and challenges. In *Findings of the Association for Computational Linguistics: ACL-IJCNLP 2021*, pages 2637–2652, Online. Association for Computational Linguistics.

Sai Surya, Abhijit Mishra, Anirban Laha, Parag Jain, and Karthik Sankaranarayanan. 2019. Unsupervised neural text simplification. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 2058–2068, Florence, Italy. Association for Computational Linguistics.

Thanh Vu, Dat Quoc Nguyen, Dai Quoc Nguyen, Mark Dras, and Mark Johnson. 2018. VnCoreNLP: A Vietnamese natural language processing toolkit. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 56–60, New Orleans, Louisiana. Association for Computational Linguistics.

Sander Wubben, Antal van den Bosch, and Emiel Krahmer. 2012. Sentence simplification by monolingual machine translation. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1015–1024, Jeju Island, Korea. Association for Computational Linguistics.

Wei Xu, Chris Callison-Burch, and Courtney Napoles. 2015. Problems in current text simplification research: New data can help. *Transactions of the Association for Computational Linguistics*, 3:283–297.

Wei Xu, Courtney Napoles, Ellie Pavlick, Quanze Chen, and Chris Callison-Burch. 2016. Optimizing statistical machine translation for text simplification. *Transactions of the Association for Computational Linguistics*, 4:401–415.

Yang Xu and David Reitter. 2016. Convergence of syntactic complexity in conversation. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 443–448, Berlin, Germany. Association for Computational Linguistics.

Xingxing Zhang and Mirella Lapata. 2017. Sentence simplification with deep reinforcement learning. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 584–594, Copenhagen, Denmark. Association for Computational Linguistics.

Zhirui Zhang, Shuo Ren, Shujie Liu, Jianyong Wang, Peng Chen, Mu Li, Ming Zhou, and Enhong Chen. 2018. Style transfer as unsupervised machine translation. *arXiv preprint arXiv:1808.07894*.

Yanbin Zhao, Lu Chen, Zhi Chen, and Kai Yu. 2020. Semi-supervised text simplification with back-translation and asymmetric denoising autoencoders. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 34, pages 05:9668–9675.

Zhemin Zhu, Delphine Bernhard, and Iryna Gurevych. 2010. A monolingual tree-based translation model for sentence simplification. In *Proceedings of the 23rd International Conference on Computational Linguistics (Coling 2010)*, pages 1353–1361, Beijing, China. Coling 2010 Organizing Committee.

A Details of Automatic Metrics

This section explains the computation of automatic evaluation metrics provided by EASSE (Alva-Manchego et al., 2019):

- **Compression ratio (CR):** average ratio of number of characters in the output to that in the original.
- **Exact Copies (CP):** proportion of outputs exactly same to the original word-wise.
- **Split ratio (%SP):** average ratio of number of sentences in the output to that in the original.
- **Additions proportion (%A):** average proportion of words in the output but not in the original.
- **Deletions proportion (%D):** average proportion of words removed from the original.
- **SARI:** for each operation $ope \in \{add, del, keep\}$, $n-$gram and its order between the output and all references, we calculate precision $p_{ope}(n)$, recall $r_{ope}(n)$ and F1 score $f_{ope}(n)$ by

$$f_{ope}(n) = \frac{2 \times p_{ope}(n) \times r_{ope}(n)}{p_{ope}(n) + r_{ope}(n)}$$

Averaging over the $n-$gram orders (k), the overall operation F1 score is

$$F_{ope} = \frac{1}{k} \sum f_{ope}(n)$$

$n = 4$ is a popular choice.

B Details of Human Evaluation

A human evaluation sheet consists of instructions followed by 50 blocks of texts. Each block contains 1 original sentence and 3 simpler variants corresponding to the simplified output from each system. The order of the variants within each block is also randomized, and we do not annotate which output belongs to which system (i.e., blind evaluation). The participants are required to use a five-point Likert scale to rate their degree of agreement to the following statements

- **Simplicity:** The output is simpler than the original sentence.
- **Adequacy:** The meaning expressed in the original sentence is preserved in the output.
- **Fluency:** The output sentence is grammatical and well formed.

You are given one original sentence and its simplified version generated by different computer programs.
Please rate each version on 3 dimensions, by indicating how much you agree with the following statements

Fluency + The output sentence is grammatical and well formed
Adequacy + The meaning expressed in the original sentence is preserved in the output
Simplicity + The output is simpler than the original sentence

1 = Strongly Disagree
2 = Disagree
3 = Neither Agree or Disagree
4 = Agree
5 = Strongly Agree

Original sentence
The Great Dark Spot is thought to represent a hole in the methane cloud deck of Neptune.

Select a value in the drop-down list

Simplifications	Fluency	Adequacy	Simplicity
1. The Great Dark Spot is supposed to represent a hole in the ice cloud deck of Neptune.	5	2	2
2. The Great Dark Spot is thought to be a hole in the cloud deck of Neptune.	5	4	4
3. The great dark spot is thought to describe a hole in the ammonia sky roof of Neptune.	4	2	1

Original sentence
He excuses himself because he has to leave for rehearsal, and he and Dr. Schön leave.

Select a value in the drop-down list

Simplifications	Fluency	Adequacy	Simplicity
1. He apologizes for having to go to trial, and he and Dr. Go nicely	2	1	3
2. He tells him that he has to leave for rehearsal, and he and Drön leave.	3	2	4
3. He excuses himself because he has to return for recital, and he and Dr. Schon return.	3	3	2

Figure 2: English Human Evaluation Interface

Dưới đây bao gồm 50 mẫu câu tiếng Việt được viết lại một cách đơn giản hơn bởi một phần mềm má
Vui lòng đánh giá mức độ đồng ý với các ý kiến dưới đây cho từng câu

Lưu loát + Câu viết lại được trình bày một cách lưu loát và trôi chảy
Đồng nghĩa + Câu viết lại gần hoặc giống với ý nghĩa của câu gốc
Đơn giản + Câu viết lại đơn giản và dễ hiểu hơn câu gốc

1 = Rất không đồng ý
2 = Không đồng ý
3 = Trung lập
4 = Đồng ý
5 = Rất đồng ý

Câu gốc
đối với các trường hợp đình chỉ giải quyết vụ án khác theo quy định của bộ luật này thì người yêu cầu xem xét, thẩm định phải chịu chi phí xem xét, thẩm định tại chỗ.

Bấm vào mũi tên để chọn

Câu viết lại	Lưu loát	Đồng nghĩa	Đơn giản
1. đối với các trường hợp đình chỉ giải quyết vụ án khác - theo quy định của bộ luật này yêu cầu xem xét , - thẩm định phải chịu chi phí xem xét - thẩm định tại chỗ .	2	4	2

Câu gốc
trường hợp tòa án xét thấy cần thiết và quyết định xem xét, thẩm định tại chỗ thì nguyên đơn, người yêu cầu giải quyết việc dân sự, người kháng cáo theo thủ tục phúc thẩm phải nộp tiền tạm ứng chi phí xem xét, thẩm định tại chỗ.

Bấm vào mũi tên để chọn

Câu viết lại	Lưu loát	Đồng nghĩa	Đơn giản
1. Trường hợp Tòa án xét thấy cần thiết và quyết định xem xét, thẩm định tại chỗ thì nguyên đơn, người yêu cầu giải quyết việc dân sự, người kháng cáo theo thủ tục phải nộp tiền tạm ứng chi phí xem xét, thẩm định tại chỗ.	3	5	3

Câu gốc
người được ghép mô, bộ phận cơ thể người có thẻ bảo hiểm y tế được cơ quan bảo hiểm y tế thanh toán viện phí về việc ghép theo quy định của pháp luật về bảo hiểm y tế.

Figure 3: Vietnamese Human Evaluation Interface

C System Effectiveness

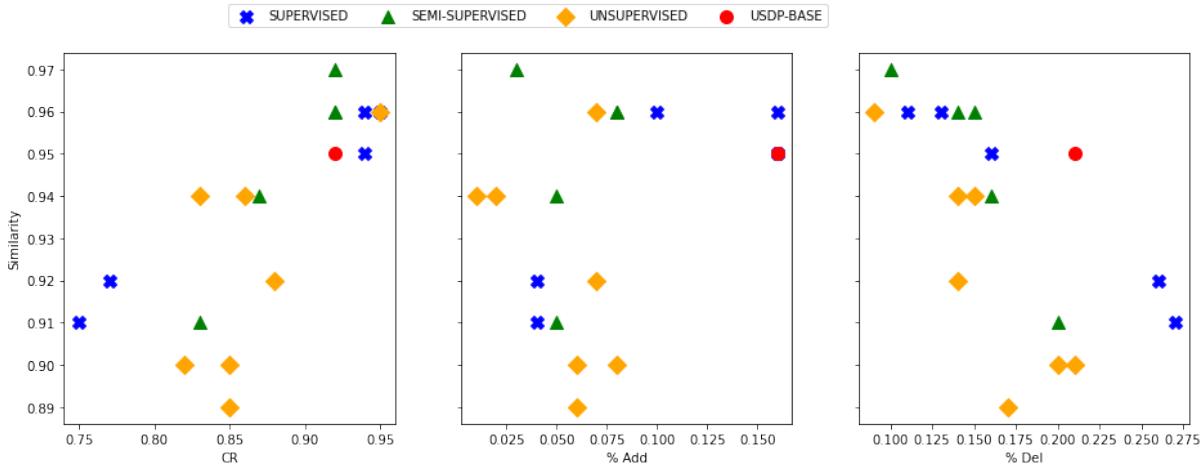


Figure 4: Visualization of systems' capacity to balance **Adequacy** with **Simplicity** on TurkCorpus

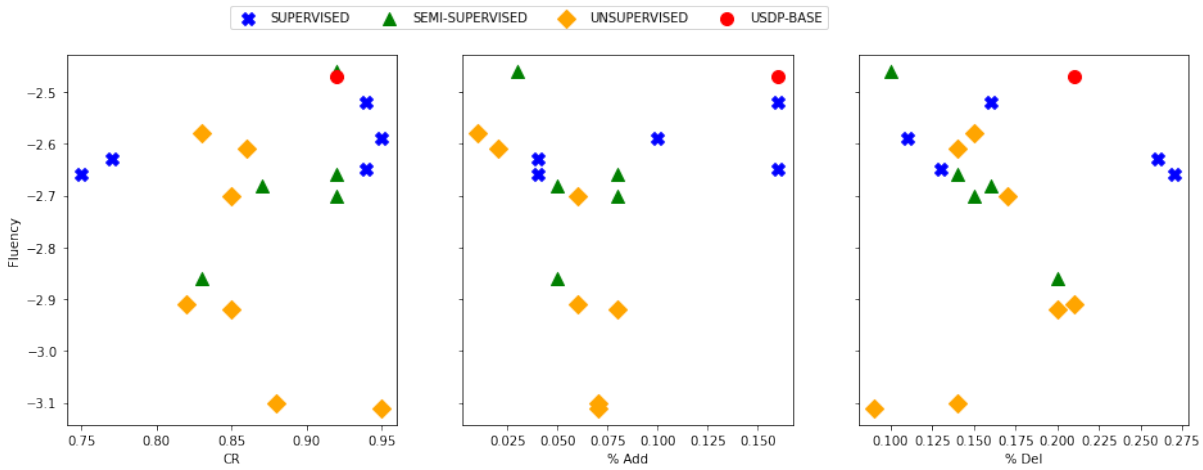
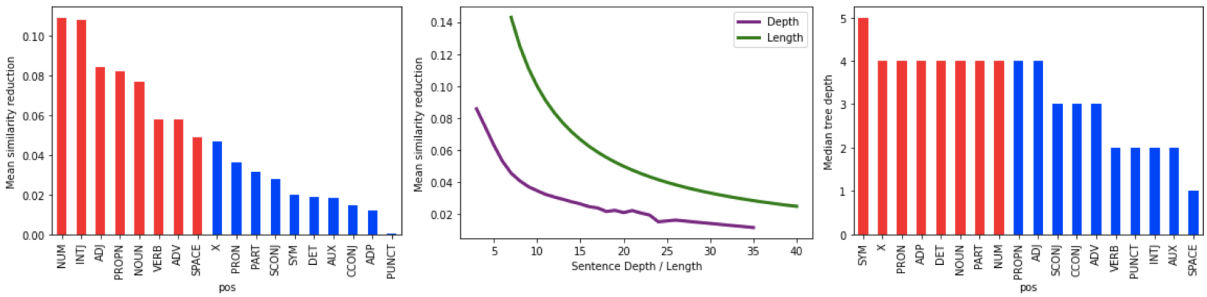


Figure 5: Visualization of systems' capacity to balance **Fluency** with **Simplicity** on TurkCorpus



(a) (b) (c)

Figure 6: Figure **(a)** shows the effect of grammatical functionality (part-of-speech) of words on reduction in similarity scores. **(b)** illustrates the correlation between location of words and reduction in similarity scores i.e., the allocation of important words in the sentence. **(c)** displays the uniform distribution of part-of-speeches across the sentence.

D Qualitative Evaluation

Example of deleting prepositional/adjective phrases	
Original sentence	*Jeddah is the **principal** gateway to Mecca, Islam's holiest city, which able-bodied Muslims **are required to visit at least once** in their lifetime.*
System outputs	
RM+EX+LS	*Jeddah is the principal gateway to Mecca, Islam's Holiest city, which Able-Bodied Muslims are required to visit at least once in their lifetime.*
USDP-Base	
With BT	*Jeddah is the **main** gateway to Mecca, the holiest city in Islam that Muslim people **had to visit** during their lifetime.*
Without BT	*Jeddah is the principal gateway to Mecca , Islam 's holiest city , which able - bodied Muslims required - visit in their lifetime.*
Example of summarization	
Original sentence	*At four-and-a-half years old he was left to fend for himself on the streets of northern Italy for the next four years, **living in various orphanages and roving through towns with groups of other homeless children.***
System outputs	
RM+ES+LS	*At Four-And-A-Half years old he was left to fend for himself on the walls of northern Italy for the next four years.*
USDP-Base	
With BT	*At the age of four and a half, he had to support himself on the streets of northern Italy for the next four years, **wandering through the cities living in various orphanages.***
Without BT	*At four - and - a - half years old he was left - to fend for himself on the streets of northern Italy for the next four years , living in various orphanages - roving through towns..*
Example of chunking	
Original sentence	*In late 2004, Suleman made headlines by cutting Howard Stern's radio show from four Citadel stations, citing Stern's frequent discussions regarding his upcoming move to Sirius Satellite Radio.*
System outputs	
RM+ES+LS	*In late 2004, Suleman made headlines by cutting Howard Stern'S radio show from four Citadel trains, reporting Stern'S serious questions.*
USDP-Base	
With BT	*In late 2004, Suleman made headlines - **by cutting Howard Stern's radio show from four Citadel stations** - **citing Stern's discussions** - **regarding the upcoming move**.*
Without BT	*In late 2004, Suleman made headlines - by cutting Howard Stern's radio show from four Citadel stations - citing Stern's discussions - regarding upcoming move.*
Example of simplistic paraphrasing	
Original sentence	*Fearing that Drek will destroy the galaxy, Clank asks Ratchet to **help him find the famous superhero Captain Qwark, in an effort to stop Drek.***
System outputs	
RM+ES+LS	*Fearing that Drek will bring the universe, Clank asks Ratchet to help him find the famous Superhero captain Qwark, in an attempt to get Drek.*
USDP-Base	
With BT	*Fearing Drek might destroy the galaxy, Clank asks Ratchet to **find the superhero in order to stop Drek**.*
Without BT	*Fearing that Drek will destroy the galaxy , Clank asks Ratchet - help find the superhero - in effort stop Drek.*

Table 7: Qualitative results on TurkCorpus.