SEMA: Text Simplification Evaluation through Semantic Alignment

Xuan Zhang[1,2], Huizhou Zhao[1,*], Kexin Zhang[1], and Yiyang Zhang[3]

[1]School of Information Science,
Beijing Language and Culture University
Email: xk18zx@126.com, zhaohuizhou@blcu.edu.cn, jacquelinelopze@outlook.com
[2]Department of Automation,
Tsinghua University
[3]College of Chinese Studies,
Beijing Language and Culture University
Email: bournezyy@163.com

Abstract

Text simplification is an important branch of natural language processing. At present, methods used to evaluate the semantic retention of text simplification are mostly based on string matching. We propose the SEMA (text Simplification Evaluation Measure through Semantic Alignment), which is based on semantic alignment. Semantic alignments include complete alignment, partial alignment and hyponymy alignment. Our experiments show that the evaluation results of SEMA have a high consistency with human evaluation for the simplified corpus of Chinese and English news texts.

1 Introduction

Text simplification is a rewriting operation that aims to improve the comprehensibility of the text by modifying, deleting, simplifying human-readable text. It tried to retain the core semantics of original text while improving readability of the text. In natural language processing tasks, long and complex sentences will bring about various problems, for example, the quality of grammatical analysis depends on the length and the grammar difficulty of texts directly, and complex sentences may cause ambiguity during machine translation(Chandrasekar and Srinivas, 1997). Therefore, text simplification is often used in the preprocessing steps of other NLP tasks. In addition, text simplification is also used to rewrite reading materials for children, second language learners, readers with aphasia and other people with low reading comprehension skills(Carroll J, 1998). As related researches are in the early stage, the results of text simplification cannot meet the needs of the audience well. One of the difficulties is the lack of reasonable text simplification evaluation indicators. At present, most evaluation methods are conducted by experts or machine translation evaluation indicators. Therefore, researches on how to analyze the results of text simplification has important application value.

Text simplification mainly includes vocabulary and semantic structure simplification. The main operation is text segmentation, that is, rewriting a single sentence into one or more simpler sentences while preserving the main semantics(Sulem et al., 2018b). Text simplification has gradually attracted attention in recent years(Xu et al., 2016; Saggion and Horacio, 2017; Saggion et al., 2012), it should be evaluated from three aspects: fluency (ie: grammatical correctness), correctness (ie: semantic retention) and simplicity (ie: degree of text simplification). Initially, experts can only evaluated the results through three aspects, and the final score is based on the Likert scale[1]; Later, someone proposed to use readability indicators to evaluate text simplification, but because the readability indicators are designed for passage-level texts, the application effects at the sentence level are not very prominent(Coster and Kauchak, 2011). In recent years, the evaluation indicators of machine translation have been increasingly used in the evaluation of text simplification, including BLEU, ROUGE based on N-gram and WER, TER based on edit distance.

In machine translation tasks, BLEU is the most widely used evaluation indicators, which was proposed in 2002. The original purpose is to replace the manual evaluation of translation results. The quality of the machine translation task is mainly evaluated by evaluating the difference between the

*Corresponding author: zhaohuizhou@blcu.edu.cn

[1]Likert scale is one of the most commonly used scoring aggregate scales. It was developed by American social psychologist Likert in 1932 on the basis of the original aggregate scale. The scale consists of a set of statements. Each statement has five answers: "strongly agree", "agree", "not necessary", "disagree" and "strongly disagree", which are recorded as 5, 4, 3, 2, 1, and the final score is the sum of score for each aspect.

Proceedings of the 6th Workshop on Natural Language Processing Techniques for Educational Applications, pages 121–128
Suzhou, China, December 4, 2020.

output generated by model and the reference. It has low computational cost and is highly correlated with human evaluation, so it is widely used. Elior Sulem's experiments show that Since the main operation of text simplification is text segmentation, involving semantic structure splitting, BLEU did not show a high degree of relevance to manual evaluation in terms of grammar and semantic retention of 70 pairs of sentences (Sulem et al., 2018c). In addition, in terms of simplicity assessment, BLEU shows a negative result which penalized simplified sentences highly.

SARI is an evaluation indicator based on reference sentences proposed in 2016(Xu et al., 2016). It focuses on the aspect of words added, deleted, and retained, but it cannot evaluate sentences at semantic level. SAMSA is a semantic structure-based evaluation indicator proposed in 2018(Sulem et al., 2018a), but it relies too much on string matching in the judgment of semantic consistency, which leads to low semantic retention calculation results for simplified text. Based on the characteristics of these evaluation indicators, this research proposes a text simplification evaluation indicator SEMA based on semantic alignment.

The contribution of this paper is to propose a semantic retention evaluation indicator of text simplification based on semantic alignment. Semantic alignment includes **complete alignment**, **partial alignment** and **hyponymy alignment**. Different semantic alignment weights are given according to the degree of semantic alignment, so as to reasonably evaluate the semantic retention of text simplification of different rewriting methods.

2 Related Work

2.1 Universal Cognitive Conceptual Annotation(UCCA)

The current traditional syntactic structure cannot directly reflect the semantic difference of the text, for example:"John took a shower." (a) and "John showered." (b) Syntactic analysis will regard them as different structures, but at the semantic level, (a) and (b) are similar. The UCCA (Universal Conceptual Cognitive Annotation)(Abend and Rappoport, 2013) proposed in 2013 avoids this defect. Its scene-based semantic structure annotation method aims to extract the scene graph formed by main relation and participants to represent the main semantic infomation in the text.

The scenes of UCCA represent motions, actions or states that persist in time, and are divided into State (S) and Process (P). A State represents a continuous state in time, such as:"There has been conflict in Syria for the last nine years." A Process describes an event that is evolving and unfolding in time, such as: "The dog runs into the house." Each scene contains a main relation, one or more participants (including location information), such as:"John kicked his ball." In this scene, the participants are "John" and "his ball", the relation is "Kicked".

The UCCA structure is a directed acyclic graph, and the smallest meaningful unit is on the leaf node (that is, the word in the text). For units that cannot form a scene, the UCCA sets a category Centers (C) to represent the subunits of a non-scene unit, and there may be one or more C in a non-scene unit. Modifiers (including qualifiers) are marked as Elaborator (E). For example, in the non-scene unit "his ball", "his" is E, and "ball" is C.

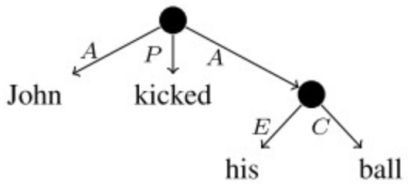

Figure 1: The result of "John kicked his ball" by UCCA

In actual contexts, more complicated situations often occure: one scene may be a participant of another scene. For example, in the sentence "The report says that the USA can be war criminals", "the USA can be war criminals" is A in the scene where "says" is a relation; one scene can also be E in another scene, such as the sentence: "The day Tom arrived in Beijing was Friday", the scene "Tom arrived in Beijing" is E that modifies "The day", and "The day" is A of the scene "The day was Friday".

UCCA is a semantic annotation method as opposed to syntactic analysis. It is portable between various fields and languages, and is not sensitive to semantic-retain grammatical changes. In addition, it can accommodate more semantic differences. In this research, the TUPA tool is used to obtain the UCCA annotation result (Hershcovich et al., 2017), it uses the NN classifier and BiLSTM model for training, inputting text and outputting UCCA result.

2.2 Simplification Automatic evaluation Measure through Semantic Annotation(SAMSA)

SAMSA is the first indicator to evaluate the quality of Text Simplification (TS) system at the semantic structure level. It uses UCCA based on the concept of scene to try to reasonably evaluate the text simplification results in terms of semantic rather than syntax(Sulem et al., 2018a). SAMSA extracts the scene of the input sentence, and after identifying the relation and participants, it does the word comparision calculation with output sentence It believes that the result of a high-quality text simplification should be: each input scene is mapped to the output sentence one by one, the smallest unit of the relation and the participants (see later) can be matched in the output sentence. SAMSA is a non-referenced automatic evaluation method. Elior Sulem's experiments show that SAMSA has a high relevance to human evaluation in terms of semantic retention. SAMSA is explained in detail below.

SAMSA is based on two external tools-UCCA and Word Alignment. UCCA decomposes each input sentence S into a set of scenes$\{\mathrm{SC}_1, \mathrm{SC}_2, \cdots \mathrm{SC}_\mathrm{n}\}$, each scene$\mathrm{SC}_i$ contains one main relationMR_i and one or more participantsA_i; Word Alignment aligns the words of the input sentence with one or zero words of the output sentence to form a set A, which can identify synonym substitution (start/begin) and stemming (run/ran). n_{inp} is the number of scenes of input, $\mathrm{n}_{out}$is the number of sentences of output (S_1, S_2 ,.., S_{nout}). Firstly, SAMSA aligns the input scene and the output sentence. There are two cases:

1.$\mathrm{n}_{inp} \geq n_{out}$: in this case, we compute the maximal Many-to-1 correspondence between Scenes and sentences. To align each input scene with the output sentence, SAMSA gets the number of word matches between each scene and each output sentence according to the word alignment A, and select the sentence with the highest matching degree to align. If $\mathrm{n}_{inp} = n_{out}$, once a sentence is matched to a scene, it cannot be matched to another one.

$$M*(SC_i) = argmax_s score\,(SC_i, S) \quad (1)$$

2.$\mathrm{n}_{inp} < n_{out}$: In this case, a scene will necessarily be split across several sentences. As this is an undesired result, SAMSA assigns this instance a score of zero.

For the scenes of input$\{\mathrm{SC}_1, \cdots \mathrm{SC}_{\mathrm{ninp}}\}$, the sentences of output$\{\mathrm{S}_1, \cdots, \mathrm{S}_{\mathrm{nout}}\}$ and their mapping relationship$M^*\,(SC_i)$, the calculation formula of SAMSA is as follows:

$$SAMSA = \begin{cases} \frac{n_{out}}{n_{inp}} \frac{1}{2n_{inp}} \sum\limits_{SC_i} \left[II_{M*(SC_i)}\,(MR_i) + \frac{1}{k_i} \sum\limits_{j=1}^{k_i} II_{M*(SC_i)}\left(Par_i^{(j)}\right) \right], & n_{inp} \geq n_{out} \\ 0, & n_{inp} < n_{out} \end{cases} \quad (2)$$

$\mathrm{MR}_i$is the smallest unit of relation in SC_i, $\mathrm{Par}_i^{(j)}\ (j = 1, \cdots, k_i)$ is the smallest unit of participants in SC_i. The smallest unit is the child node marked as C in the UCCA graph starting recurrence from P/S and A until the leaf node. If the participant is a scene, its smallest unit is the main relation of the scene. For example, the center of "the tallest building in the world" (u1) is "the tallest building". The center of the latter is "building", which is a leaf node. Therefore, the smallest unit of u1 is "building".

$\mathrm{II}_s\,(u)$ defines a function with a value between 0 and 1. If there is a word alignment in u and s, the value is 1, otherwise the value is 0. SAMSA sets a penalty factor $\mathrm{n}_{out}/\mathrm{n}_{inp}$ to penalize the case of $\mathrm{n}_{inp} > n_{out}$. In addition, SAMSA-abl is also set as the calculation indicator for removing the penalty coefficient, and the calculation is shown in formula 3. Elior Sulem's experiment (Sulem et al., 2018a) shows that the evaluation result of the SAMSA-abl indicator (0.54), which removes the penalty coefficient, is better than SAMSA. It indicates that the penalty coefficient will over-punish the situation of $\mathrm{n}_{inp} > n_{out}$, so this research improves the indicator based on SAMSA-abl.

$$SAMSA = \begin{cases} \frac{1}{2n_{inp}} \sum\limits_{SC_i} \left[II_{M*(SC_i)}\,(MR_i) + \frac{1}{k_i} \sum\limits_{j=1}^{k_i} II_{M*(SC_i)}\left(Par_i^{(j)}\right) \right], & n_{inp} \geq n_{out} \\ 0, & n_{inp} < n_{out} \end{cases} \quad (3)$$

To make the calculation process of SAMSA-abl clearer, we take the input sentence (a) "About 13 million Syrians had to leave their homes because of danger." and the simplified sentence (b) "About 13 million had to leave their homes." as an example. The smallest unit of the main relation of input scene is "leave", and the smallest unit of participants is "About, 13, million", "Syrians" and "homes". In all the smallest units, only "Syrians" in the simplified sentence fails to match the input sentence. Therefore, $\mathrm{II}_{M*(sc_1)}\,(MR_1)$ is 1, $\mathrm{II}_{M*(sc_1)}\,(Par_i)$ is 1+0+1=2(k=3), and the score of (b) is 1/2* （1+1/3*2） =0.83.

3 Text Simplification Evaluation Through Semantic Alignment(SEMA)

SEMA is a further optimization of the SAMSA indicator, including two parts: 1. The basic for-

mula SEMA-base (basic formula) is obtained by calculation when $\mathrm{n}_{inp} < n_{out}$ is added on the basis of SAMSA-abl; 2. In terms of indicator calculation strategy, semantic alignment is used to replace the string alignment and it mainly includes three semantic alignment methods: full alignment (SEMA-base), partial alignment, and hyponymy alignment.

3.1 SEMA-base

SAMSA believes that when $\mathrm{n}_{inp} < n_{out}$, a scene is broken into multiple sentences, which destroyes the structure of the scene, so the score is 0. However, in the corpus used in this research, there are more texts that meet $\mathrm{n}_{inp} < n_{out}$. For example, in the original sentence "Central Park Tower has just become the tallest residential building in the world", the simplified text is divided into four sentences:" (1)Central Park Tower is a building in New York. (2)There are only apartments in this building. (3)There are no offices in this building. (4)Now, it is the tallest building with apartments in the world." Although this text divides a scene into multiple sentences, from the perspective of reading comprehension, the simplified sentence is easier to understand and also retains the semantics of original sentence. It is unreasonable to get 0 under the condition of $\mathrm{n}_{inp} < n_{out}$.

Based on this point, on the basis of SAMSA-abl, the definition of SEMA-base is shown in formula 4, where when $\mathrm{n}_{inp} < n_{out}$, the simplified text can still get a score.

$$SEMA-base = \frac{1}{2n_{inp}} \sum_{SC_i} \left[II_{M*(SC_i)}(MR_i) + \frac{1}{k_i} \sum_{j=1}^{k_i} II_{M*(SC_i)(Par^{(j)})} \right] \quad (4)$$

3.2 Computing Strategy changes

SAMSA relies too much on string-match when aligning the words of the input scene and the output sentence, which leads to low evaluation results easily. SEMA changes the calculation and matching method based on SEMA-base, and emphasizes semantic alignment, including complete alignment, partial alignment and hyponymy alignment. Complete alignment is the original SAMSA string-match strategy.

Partial Alignment: SAMSA requires that the smallest unit of the participant in the scene matches the word of the output sentence. For the case where a participant contains multiple smallest units, SAMSA requires that all smallest units should be matched to get score 1, otherwise it is 0. For example, for the input sentence "I like banana, apple and orange.", the participants are "banana, apple, orange". When the output sentence is "I love apple.", only "apple" is matched in the smallest unit of the participant, but the value is 0 according to the SAMSA matching method. Obviously, this is not friendly to sentences that contain part of the smallest unit. Partial alignment calculates the matching degree of every single smallest unit and SEMA-part is defined as shown in formula 5. On the basis of SEMA-base, the parameter m_q is added to represent the number of smallest units of participants, and $Par_i^{(j)(q)}$ is the qth smallest unit of the jth participant in the ith scene.

$$SEMA-part = \frac{1}{2n_{inp}} \sum_{SC_i} \left[II_{M*(SC_i)}(MR_i) + \frac{1}{k_i} \sum_{j=1}^{k_i} \frac{1}{m_q} \sum_{q=1}^{m_q} II_{M*(SC_i)\left(Par_i^{(j)(q)}\right)} \right] \quad (5)$$

Hyponymy Alignment: In order to summarize the text features of text simplification better and establish a more complete evaluation indicator in terms of semantic evaluation, we observed and disassembled the corpus, compared the manual score with automatic machine score, and found the feature of hyponymy in the corpus. It is a common operation to replace hyponym with hypernym in text simplification. Here, the hyponymy refers to the words with the upper and lower conceptual relationship, and they have a species relationship (Chi, 1989), such as "drinks" is the hypernym of "beer", "fruit" is the hypernym of "kiwi". Generally, the most simplified text has more hyponymy. In this research, based on SEMA-part, we use WordNet's hyponymic relationship network to align the smallest unit of relations and participants which include hyponymy. And it improves the degree alignment between the input scene and the output sentence. Finally, a text simplification evaluation indicator based on semantic alignment SEMA is formed. The calculation formula of SEMA is still shown in formula 5. The difference between SEMA-part and SEMA is only the addition of hyponymy alignment to the semantic alignment. In the end, our experiments proved that SEMA is highly usable in evaluating the semantic retention of Chinese and English text simplification at sentence and passage level. See chapter 4 for more details.

4 Evaluation Experiment Based On Artificial Simplified Corpus

4.1 Corpus

This research uses simplified Chinese and English news corpus for experiments. The simplified

English corpus comes from the English website: News in Levels, which is a free online news website specially designed for English students. Each article is written in three levels, and level 1 is the simplest. Taking level 3 as the benchmark, the semantic retention of level 2 and level 1 is manually judged to be around 70% and 50% respectively. The Chinese news corpus comes from the texts of the Chinese news reading textbook and its corresponding original texts. The texts of the news reading textbook are simplified and adapted for teaching needs. The semantic retention of the adapted text is around 80%. This research collected 200 English passages (three levels), 600 pieces in total, 100 aligned sentences; 100 Chinese aligned sentences.

4.2 English Corpus Experiment

We first perform experiments on SAMSA, SAMSA-abl, SEMA-base, SEMA-part, and SEMA on 100 English sentences. The experimental results are shown in Table 1.

sentence level		
	level1	level2
SAMSA	0.22	0.36
SAMSA-ab1	0.34	0.62
SEMA-base	0.43	0.65
SEMA-part	0.45	0.67
SEMA	0.48	0.69

Table 1: Sentence-level results of SAMSA and SEMA

The results show that SAMSA does not evaluate the semantic retention of each level of corpus very well; after removing the penalty coefficient, SAMSA-abl significantly improves the scores of the two levels. It proves that the penalty coefficient will over-punish the corpus; when considering the case: $\mathrm{n}_{inp} < n_{out}$,the scores of level1 and level2 are improved and the degree of improvement of level1 is more obvious, which also matches the corpus characteristics of level1 (more corpus conforms to $\mathrm{n}_{inp} < n_{out}$); After adding partial alignment and hyponymy alignment, the results of the corpus evaluated by SEMA are closer to the human estimated scores, with level1-score increased to 0.48 and level2-score increased to 0.69. The effect of each optimization strategy on the experimental results is shown in Figure 2.

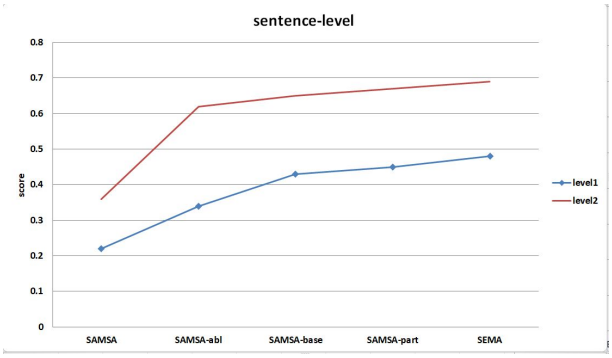

Figure 2: Sentence-level semantic retention evaluation, the improved experimental results of each optimization strategy

SAMSA is proposed to evaluate the sentence-level text simplification system. This research applies it to the passage-level evaluation. Firstly, 35 passages corresponding to 100 alignment sentences are selected for experiment. Based on SEMA-base, the result of level1 is 0.38, and the result of level2 is 0.52.

It can be seen from the experimental results that the overall score of the passage level is lower than the sentence level. This is because in the passage-level evaluation, the length of the passage and the sentence number increase, the scene analysis tool TUPA is unstable. Therefore, it is difficult to extract the scene (that is, multiple sentences extract a large scene), such as the sentence "They based the report on hundreds of interviews and analyses of photos, videos, and satellite images." should be divided into one scene, but in the actual results, "videos, and satellite images" and "Put simply" which is far away are seen as one scene. Since the input scene and the output sentence are aligned according to the maximum number of word matches, and the scene cannot be split clearly, multiple scenes can only be aligned with one sentence. Obviously, it is difficult to find all the semantic information of multiple scenes in one sentence in this case, which directly affects the quality of the indicator evaluation. In order to improve this shortcoming, we splitted the original passages (level3) and then used TUPA for scene analysis. The scene analysis result of each sentence was compared with the simplified whole passage, so the best match can be selected. The final score is averaged.

The experimental results at the passage level are shown in Table 2. "Segmentation+SEMA-base" is an improvement based on SEMA-base. It can be

seen that the division of the passage helps TUPA extract the scene and improve the accuracy of the indicator. In the end, the evaluation results of 35 passages of level1 and level2 increased from the initial 0.24 and 0.26 to 0.53 and 0.69 respectively. When we expand the corpus from 35 passages to 200 passages, level1 and level2 scores are 0.51 and 0.68 respectively, which is consistent with the manual evaluation results.

	passage level	
	level1	level2
SEMA-base	0.38	0.52
Segmentation+SEMA-base	0.44	0.62
Segmentation+SEMA-part	0.45	0.64
Segmentation+SEMA	0.53	0.69

Table 2: Passage-level results of SEMA

As for the passage level, the effect of each indicator optimization strategy on the experimental results is shown in Figure 3. Among them, the improvement of hyponymy alignment is obvious, and the performance on level 1 is particularly prominent. In the final SEMA evaluation results, the passage level score of level 1 is much higher than the sentence level. The main reason is that when we align the sentences, we filter out some improperly aligned sentences, and all sentences at the passgae level participate in the scoring. The scores of these sentences increase the average score at the passage level.

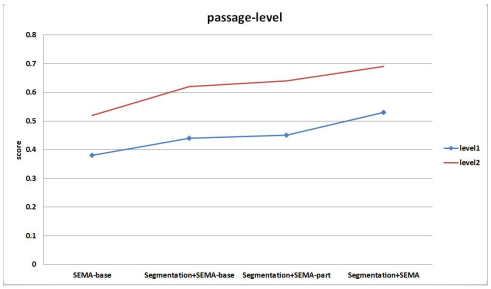

Figure 3: Passage-level semantic retention evaluation, the improved experimental results of each optimization strategy

4.3 Chinese Corpus Experiment

Compared with English, Chinese is consistent with English in the main output sequence of sentences such as subject, predicate and object. For some subsidiary components, such as attributes that modify the subject and object, there are many differences between Chinese and English. The Chinese corpus comes from the adapted Chinese news reading textbook and its corresponding original text. The adaptation methods include but are not limited to: deletion, replacement, and rewriting. According to manual evaluation, the semantic retention of the adapted corpus is about 80% or more.

This research uses semi-automatic processing in the Chinese corpus experiment. There are no tool to analyse Chinese UCCA structure, when extracting the main information of the input sentence, we use Baidu dependency syntax analysis [2] to extract the core word of the sentence (HED) as the main relation, the first-level child nodes of the core words are participants. For example, in the sentence "2004年3月26日全法汉语教学研讨会在巴黎国际大学生城举行。", the relation is "举行", the participants are "研讨会" and "在巴黎". For semantic alignment, we use manual alignment in a non-automated way, and finally conduct experiments based on aligned 100 Chinese sentences. The results are shown in Table 3.

	Adapted text
SEMA	0.804

Table 3: SEMA evaluation results at the Chinese sentence level

Experiments show that in evaluating the semantic retention of the adapted Chinese sentences, SEMA reaches to 0.804, which is consistent with the manual evaluation result. This has great significance for the evaluation of the semantic retention of Chinese text simplification.

5 Conclusion

This research improves the semantic structure based text simplification evaluation measure SAMSA proposed in 2018. There are mainly several aspects: the case of $\mathrm{n}_{inp} < n_{out}$ is considered on the basis of SAMSA-abl; semantic alignment is used to replace string matching, mainly based on three semantic alignments method: Full alignment, partial alignment, hyponymy alignment. Finally, a semantic retention evaluation measure about text simplification SEMA based on semantic alignment

[2]The dependency syntax explains its syntactic structure by analyzing the dependencies of the components in the language unit, claiming that the core verb in the sentence is the central component that dominates other components

is formed. We did experiments on English sentence-level and passage-level. The experimental results show that it is similar to the manual evaluation results, which shows its significance in text simplification evaluation.

When we apply SEMA to Chinese, we summarize the characteristics of Chinese and use dependency syntax analysis to extract the main semantic information in the sentence. Experimental results show that SEMA has high applicability in Chinese corpus, and it is the first semantic retention evaluation indicator based on semantic alignment on Chinese corpus.

In future research, we will continue to use larger corpus to explore SEMA's evaluation methods under different semantic retention thresholds; In addition, the text simplification indicator proposed in this paper only evaluates the semantic retention, other aspects of evaluating texts simplification such as grammaticality and degree of simplification need to be further explored; At the same time, follow-up research should expand the scale of the text corpus and collect multi-subject, multi-genre and multi-length texts to test the usability of our indicator.

Acknowledgments

The research is supported by Science Foundation of Beijing Language and Culture University (supported by "the Fundamental Research Funds for the Central Universities")(19YJ040005); Major Program of National Social Science Foundation of China (18ZDA295); Top-ranking Discipline Team Support Program of Beijing Language and Culture University(JC201902).

References

Omri Abend and Ari Rappoport. 2013. Universal conceptual cognitive annotation (ucca). In *Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*.

Xiaohong Yuan Guodong Zhou Bukang Wang, Hongling Wang. 2010. Chinese semantic role tagging based on dependency syntax analysis. *Journal of Chinese Information Processing — J Chin Inf Proc*, 01:25–29. (in Chinese).

Canning Y Devlin S Tait J Carroll J, Minnen G. 1998. Practical simplification of English newspaper text to assist aphasic readers. In *Proceedings of the AAAI-98 Workshop on Integrating Artificial Intelligence and Assistive Technology*, pages 7–10.

R. Chandrasekar and B. Srinivas. 1997. Automatic induction of rules for text simplification. *Knowledge Based Systems*, 10(3):183–190.

Mei Chi. 1989. Talking about hyponymy. *HAN YU XUE XI*, (01):26–28. (in Chinese).

Will Coster and David Kauchak. 2011. Learning to simplify sentences using wikipedia. In *Proceedings of the Workshop on Monolingual Text-To-Text Generation*.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*.

Daniel Hershcovich, Omri Abend, and Ari Rappoport. 2017. A transition-based directed acyclic graph parser for ucca. *arXiv preprint arXiv:1704.00552*.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

George Miller. 1995. Wordnet: A lexical database for English. *Communications of the ACM*, 38:39–.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting of the Association for Computational Linguistics*, pages 311–318.

Saggion and Horacio. 2017. Automatic text simplification. *Synthesis Lectures on Human Language Technologies*, 10(1):1–137.

Horacio Saggion, Elena Gómezmartínez, Esteban Etayo, Alberto Anula, and Lorena Bourg. 2012. Text simplification in simplext. making text more accessible. In *International Conference on Computational Science*.

Matthew Snover, Bonnie Dorr, Richard Schwartz, Linnea Micciulla, and John Makhoul. 2006. A study of translation edit rate with targeted human annotation. In *Proceedings of association for machine translation in the Americas*, volume 200. Cambridge, MA.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018a. Semantic structural evaluation for text simplification. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018b. Simple and effective text simplification using semantic and neural methods. In *Meeting of the Association for Computational Linguistics*.

Elior Sulem, Omri Abend, and Ari Rappoport. 2018c. BLEU is not suitable for the evaluation of text simplification. *arXiv preprint arXiv:1810.05995*.

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 Lingus*, 4(4):401–415.