SMSMix: Sense-Maintained Sentence Mixup for Word Sense Disambiguation

Hee Suk Yoon[1]*, Eunseop Yoon[1]*, John Harvill[2], Sunjae Yoon[1], Mark Hasegawa-Johnson[2], and Chang D. Yoo[1]†

[1]Korea Advanced Institute of Science and Technology (KAIST)

[2]University of Illinois at Urbana-Champaign (UIUC)

{hskyoon, esyoon97, sunjae.yoon, cd_yoo}@kaist.ac.kr

{harvill2, jhasegaw}@illinois.edu

Abstract

Word Sense Disambiguation (WSD) is an NLP task aimed at determining the correct sense of a word in a sentence from discrete sense choices. Although current systems have attained unprecedented performances for such tasks, the nonuniform distribution of word senses during training generally results in systems performing poorly on rare senses. To this end, we consider data augmentation to increase the frequency of these least frequent senses (LFS) to reduce the distributional bias of senses during training. We propose Sense-Maintained Sentence Mixup (SMSMix), a novel word-level mixup method that maintains the sense of a target word. SMSMix smoothly blends two sentences using mask prediction while preserving the relevant span determined by saliency scores to maintain a specific word's sense. To the best of our knowledge, this is the first attempt to apply mixup in NLP while preserving the meaning of a specific word. With extensive experiments, we validate that our augmentation method can effectively give more information about rare senses during training with maintained target sense label.

1 Introduction

Determining the meaning of a word in a particular sentence is a fundamental problem in Natural Language Processing (NLP), as it can enable a better understanding of natural languages and help solve various NLP problems. Word Sense Disambiguation (WSD) is a crucial task towards accomplishing this goal, where it involves choosing the most relevant meaning of a target word in context from predefined sense labels. Like most other NLP tasks, the advancement of Deep Learning has led supervised learning of neural models to be the primary method of WSD (Huang et al., 2019; Blevins and Zettlemoyer, 2020; Barba et al., 2021). However, one of the biggest challenges of WSD is overcoming the data bias that naturally stems from the distributional bias of senses in language (Kilgarriff, 2004). Because the dataset tends to have this bias toward certain senses, the WSD system shows high accuracy on the most frequent sense (MFS) and low accuracy on the least frequent sense (LFS) of a word.

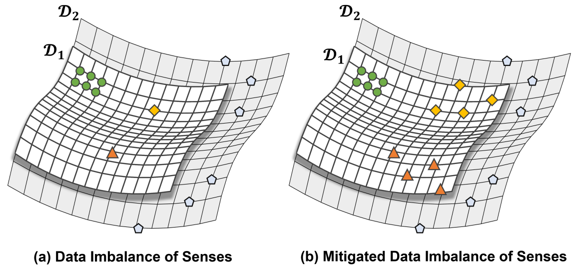


Figure 1: Schematic illustration of SMSMix filling underpopulated areas in the training data space ($\mathcal{D}_1$) - linked to least frequent senses (LFS) - with synthetic examples created through mixup with external corpus ($\mathcal{D}_2$) sentences.

A common solution in machine learning to address data imbalance involves oversampling underrepresented categories, although merely doing so often leads to overfitting of minority classes (Chawla et al., 2002). Another approach uses data augmentation to drive the learning process toward a more suitable solution. Due to the expensive cost of data collection, data augmentation has become one of the essential tools in modern deep learning, especially when dealing with low-resource tasks. In image classification, some approaches (Chou et al., 2020; Kabra et al., 2020; Galdran et al., 2021) have been recently explored, involving the well-known data augmentation method Mixup (Zhang et al., 2017) to alleviate data imbalance. This motivation encourages us to fill underpopulated areas in the training data of WSD - linked to least frequent senses (LFS) - with synthetic examples created through mixup (Figure 1).

*Equal contribution

†Corresponding author

Findings of the Association for Computational Linguistics: EMNLP 2022, pages 1493–1502
December 7-11, 2022 ©2022 Association for Computational Linguistics

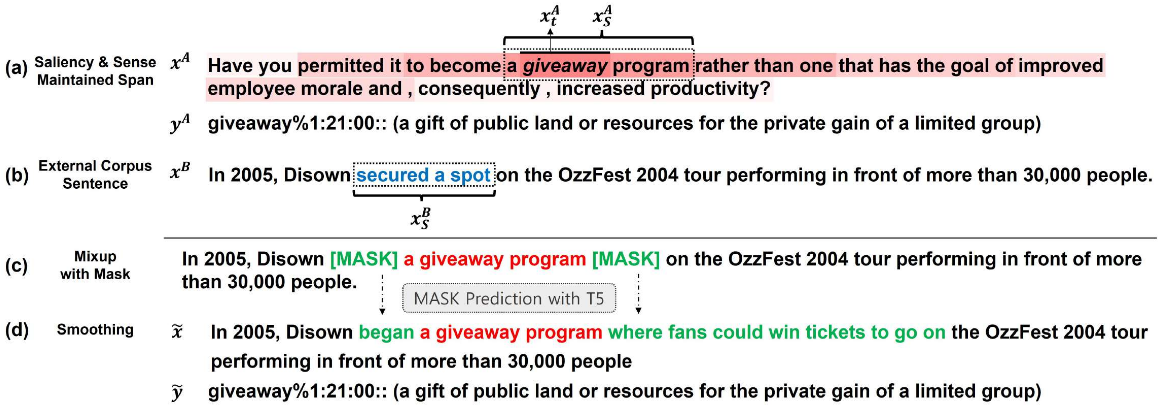


Figure 2: Overview of SMSMix. (a) x^A is a sentence containing target word x_t^A with sense label y^A. The saliency score of each token is visualized by the concentration of the color. The sense-maintained span x_S^A is determined by setting the highest saliency token from the front and back of the target word x_t^A as the start and end of the span, respectively. (b) A random sentence x^B with random span x_S^B is sampled from an external corpus. (c) x_S^A is replaced into x_S^B with mask tokens. (d) Mask token prediction with T5 smoothly blends the two sentences to minimize the strong-edge problem. The output results in $\tilde{x}$ containing the target word x_t^A with label $\tilde{y}$ set to y^A.

In this work, we propose Sense-Maintained Sentence Mixup (SMSMix), a novel mixup data augmentation method where the meaning of a specific word in a sentence is unaffected as much as possible while exposing it to various contexts. Inspired by the recently proposed word-level mixup in NLP for text classification tasks (Yoon et al., 2021), we first determine a span of text in a sentence containing a target word that is most relevant to maintaining the target word sense using gradient-based saliency scores. Second, because Wu et al. (2020) mentions that mixup within training data cannot give more information to the model and acts merely as a regularizer, we inject the aforementioned span of text into a random sentence from a Wikipedia corpus. Lastly, when carrying out this injection, we consider smoothly blending the span to the Wikipedia corpus sentence using mask prediction, inspired by SmoothMix (Lee et al., 2020), where they consider smoothly mixing two images to solve the strong-edge problem.

SMSMix has empirically shown to be an effective augmentation method that can give more information about rare senses during training a WSD model in a supervised setting. We especially show that performing mixup with an external sentence (i.e., a sentence that is not from the training data) can outperform internal mixup within training data. Also, by visualizing the latent vector of the target words in the augmented sentences, we show that SMSMix effectively maintains the sense of the target word.

2 SMSMix

SMSMix generates a new sentence $\tilde{x}$ by injecting span x_S^A, which contains a target word x_t^A with sense label y^A, from sentence x^A into another span x_S^B from sentence x^B. Prior to injection, we concatenate mask tokens to the front and back of span x_S^A and perform mask prediction to smoothly blend the two sentences. Because we use saliency score to preserve the sense label when determining span x_S^A, we set the label of x_t^A in the new sentence $\tilde{x}$ to y^A (Figure 2).

2.1 Saliency and Sense-Maintained Span

Saliency, which shows how each input fraction affects the final prediction, is usually measured using gradient-based methods. Yoon et al. (2021) recently proposed a mixup strategy in NLP using a gradient-based saliency score to preserve the locality of the two sentences performing mixup. Similarly, we compute the gradient of classification loss L with respect to the input token embedding e, and apply the L2 norm to get the saliency for each input token: i.e., $s = ||\frac{\partial L}{\partial e}||_2$. The saliency of each input token signifies how influential each input token is for the meaning of the target word. The most salient span for maintaining the target word sense is determined by taking the token with the highest saliency score at the front and back of the target word and setting it to the beginning and the end of the span, respectively.

2.2 Mixing Sentence

We consider two scenarios when injecting the sense-maintained span. First, we consider injecting the span into a random MFS sentence from the training set for the corresponding sense (internal). In this case, x_S^B is determined in the same way as x_S^A. However, Wu et al. (2020) mentions that mixup inside the training data has limitations in giving new information to the model and is effective only as a regularization role. Thus, we also consider injecting the span into a random sentence from the Wikipedia corpus (external).

2.3 Smoothing

SmoothMix (Lee et al., 2020) has been shown to be an effective mixup strategy in image classification for minimizing the problem of unnatural strong-edge of the boundary between two images performing the mixup. Motivated by SmoothMix, we propose to minimize the strong-edge problem in the boundary between the two contributing sentences. Prior to injecting the sense-maintained span onto the other sentence, we put a mask token at the front and back of the span. We then use T5 [1] (Raffel et al., 2019) for mask prediction to generate a smoothly transitioning mask and use this to blend two different texts to form an augmented sample.

3 Experimental Setup

3.1 Dataset and Model

Dataset We use the SemCor (Miller et al., 1993) for training, the largest dataset manually annotated with sense from WordNet that contains 226,036 annotated examples covering 33,362 unique senses. As standard procedure, we use SemEval-2007 (**SE07**, Pradhan et al. (2007)) as the validation set, while performing testing on Senseval-2 (**SE2**; Palmer et al. (2001)), Senseval-3 (**SE3**, Snyder and Palmer (2004)), SemEval-2013 (**SE13**, Navigli et al. (2013)), and SemEval-2015 (**SE15**, Moro and Navigli (2015)). Additionally, to measure the degree to which the system generalizes to LFS, unseen words, and definitions (zero-shot settings), we consider five subsets of the data (MFS, LFS, 0-lex, 0-lex-def, 0-def) proposed in Barba et al. (2021).

Model We evaluate our augmentation scheme on the BEM (Blevins and Zettlemoyer, 2020) system, which employs a bi-encoder to represent the target word and its sense definitions within the same space.

3.2 Training

We adopt a two-stage training strategy as Yoon et al. (2021) and Liu et al. (2021). After fully training the WSD system, we train one additional epoch with data augmentation with a learning rate of 5e-7. All other hyperparameters are set equal to the original BEM training [2]. We consider three types of data augmentation: (1) simple oversampling of LFS (Oversample), (2) SMSMix with internal training data sentence (SMSMix$_{int}$), (3) SMSMix with random external Wikipedia corpus sentence (SMSMix$_{ext}$). Additionally, for both settings of SMSMix, we automatically filter out those augmented sentences that are not grammatically acceptable, which could happen when the T5 could not smoothly blend the span into a sentence [3]. For all three data augmentations, we add three data samples for half of the LFS in the SemCor training data. When training with data augmentation, we concatenate the original data with the augmented data to prevent the training data distribution from getting too far from the original data distribution, as mentioned in He et al. (2019). All training was done on NVIDIA Quadro RTX 8000.

3.3 Evaluation

Performance of the WSD task has generally been reported so far in terms of micro-average F1 scores. However, doing so gives more weight to frequent senses simply because they occur more often [4], thus resulting in an underrepresentation of the low performances of the least frequent senses. Therefore, in addition to the micro-averaged F1 scores, we also choose to report the macro-averaged F1 scores as Maru et al. (2022).

4 Results and Discussion

4.1 Overall Results

Table 1 shows the overall micro (m-F1) and macro F1 (M-F1) results on the English all-words WSD task. We find that oversampling and SMSMix within training data (SMSMix$_{int}$) have a similar slight increase over the original BEM. Per-

[1] We use T5 Version 1.1-large for all the mask predictions. https://huggingface.co/google/t5-v1_1-large

[2] https://github.com/facebookresearch/wsd-biencoders

[3] We use the sentence acceptability judgment capability of T5-large using the "cola sentence:" prompt.

[4] Out of 6,368 test data in the aggregated ALL evaluation set, 4949 are MFS and 1419 are LFS.

Methods	Dev Set		Test Sets									
	SE07		SE2		SE3		SE13		SE15		ALL	
	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1
BEM	**74.5**	**72.4**	79.4	75.5	77.4	73.0	79.7	77.0	81.7	79.1	79.0	73.9
+Oversample	74.1	72.0	79.6	75.8	77.7	73.5	79.7	**77.7**	81.8	79.4	79.1	74.6
+SMSMix$_{int}$	74.3	72.0	79.4	75.7	**77.9**	**73.8**	79.3	77.5	81.9	79.1	79.1	74.4
+SMSMix$_{ext}$	**74.5**	**72.4**	**79.9**	**76.4**	77.8	**73.8**	**79.8**	**77.7**	**82.2**	**79.6**	**79.3**	**74.8**

Table 1: Comparison of both micro (m-F1) and macro (M-F1) F1 scores on the all-words WSD task against its baseline (BEM). The reported values are an average over five runs with different seeds (The standard deviations are reported in Appendix 5). Best macro and micro F1 scores for each test set are in **bold**.

	MFS		LFS		0-lex		0-lex-def		0-def	
	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1
BEM	**89.3**	79.0	53.3	57.7	90.2	90.3	68.2	65.7	69.1	67.1
+ Oversample	89.0	79.3	54.7	59.0	90.7	90.9	68.8	66.7	**69.7**	68.0
+ SMSMix$_{int}$	89.2	79.2	54.6	59.3	90.5	90.9	68	66.3	68.2	67.2
+ SMSMix$_{ext}$	**89.3**	**79.7**	**55.0**	**59.7**	**91**	**91.6**	**69.0**	**66.9**	**69.7**	**68.2**

Table 2: Comparison of both micro (m-F1) and macro (M-F1) F1 scores on MFS, LFS, and zero-shot datasets. The reported values are an average over five runs with different seeds (The standard deviations are reported in Appendix 6). Best macro and micro F1 scores for each dataset are in **bold**.

forming SMSMix with external Wikipedia corpus (SMSMix$_{ext}$) shows the highest increase in performance, obtaining 79.3 m-F1 and 74.8 M-F1 on the aggregated ALL evaluation set, outperforming the original BEM by 0.3 m-F1 and 0.9 M-F1 points, respectively.

4.2 Results on Sense Frequency

In Table 2 we report the results on the five subsets of the data (MFS, LFS, 0-lex, 0-lex-def, 0-def). Compared with BEM without data augmentation, the performance on LFS of oversampling, SMSMix$_{int}$, and SMSMix$_{ext}$ improved by 1.3, 1.6, and 2.0 M-F1, respectively. As the performance is improved in all three cases, it can be seen that the approach to mitigate sense imbalance through data augmentation is effective. In addition, the results have shown that performing mixup augmentation through external information obtained more performance gain than the internal mixup.

4.3 Sense-Maintained Augmentation

We visually analyze whether SMSMix changes the meaning of the target word. We first take the context encoder of a BEM system fully trained on the SemCor dataset without augmentation. Then, we apply SMSMix to the SemCor training set to generate 50 augmented sentences per sense. Moreover, we obtain 50 labeled sentences per sense from the OMSTI dataset (Taghipour and Ng, 2015). These are then fed into the pre-trained BEM system, and we extract the embedding of the target words. We plot the 2-D representation of these embeddings using t-SNE (Van Der Maaten, 2014). We find that the resulting latent space visualization of the target words in augmented sentences closely overlaps

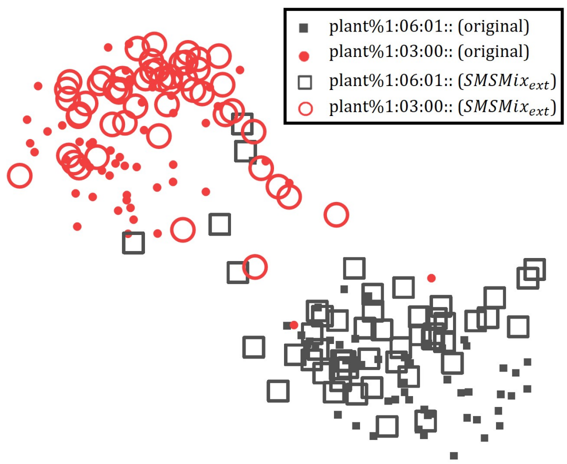


Figure 3: Latent space visualization of the embeddings of target word *plant (noun)* in the labeled OMSTI dataset sentences (original) and augmented sentences using SMSMix on SemCor training dataset. The embeddings of SMSMix target words closely overlap with those of original target words, suggesting that SMSMix maintains the true sense label of the target word.

with those in labeled sentences (Figure 3), which suggests that SMSMix effectively preserves the meaning of the target word while mixing it into various contexts.

5 Conclusion

This paper introduced a novel input-level mixup data augmentation scheme SMSMix for improving the Least Frequent Sense (LFS) data imbalance in the Word Sense Disambiguation task. SMSMix maintains the meaning of a specific word in a sentence by keeping the sense-maintained span using saliency score and smoothly injects the span into a different context using mask prediction. Throughout the experiment, we show that instead of injecting the sense-maintained span with an internal training data sentence, injecting it into a random external corpus sentence allows the model to better improve the performance on LFS.

Limitations

In this paper, we considered word-level mixup data augmentation to create new synthetic sentences containing a specific word with preserved meaning. We show in Appendix A.2 that when the two sentences performing the mixup have a high contextual difference, the T5 model fails to smoothly blend the two sentences during mask prediction, resulting in a sentence that is grammatically incorrect or does not make sense. For future work, we plan on considering sentence similarity to choose sentences for mixup instead of random selection, as in the paper.

Ethics Statement

This paper does not violate the use of others' work without reference. Furthermore, the paper does not involve introducing new datasets and the experiments conducted do not utilize demographic or identity characteristics.

Acknowledgements

This work was supported by Institute of Information & communications Technology Planning & Evaluation (IITP) grant funded by the Korea government(MSIT) (No. 2022-0-00951, Development of Uncertainty-Aware Agents Learning by Asking Questions), and was partly supported by the National Research Foundation of Korea (NRF) grant funded by the Korea government(MSIT) (No. 2022R1A2C201270611).

References

Edoardo Barba, Tommaso Pasini, and Roberto Navigli. 2021. ESC: Redesigning WSD with extractive sense comprehension. In *Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 4661–4672, Online. Association for Computational Linguistics.

Terra Blevins and Luke Zettlemoyer. 2020. Moving down the long tail of word sense disambiguation with gloss informed bi-encoders. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1006–1017, Online. Association for Computational Linguistics.

N. V. Chawla, K. W. Bowyer, L. O. Hall, and W. P. Kegelmeyer. 2002. SMOTE: Synthetic minority oversampling technique. *Journal of Artificial Intelligence Research*, 16:321–357.

Hsin-Ping Chou, Shih-Chieh Chang, Jia-Yu Pan, Wei Wei, and Da-Cheng Juan. 2020. Remix: Rebalanced mixup. *CoRR*, abs/2007.03943.

Adrian Galdran, Gustavo Carneiro, and Miguel Ángel González Ballester. 2021. Balanced-mixup for highly imbalanced medical image classification. *CoRR*, abs/2109.09850.

Zhuoxun He, Lingxi Xie, Xin Chen, Ya Zhang, Yanfeng Wang, and Qi Tian. 2019. Data augmentation revisited: Rethinking the distribution gap between clean and augmented data. *CoRR*, abs/1909.09148.

Luyao Huang, Chi Sun, Xipeng Qiu, and Xuanjing Huang. 2019. Glossbert: BERT for word sense disambiguation with gloss knowledge. *CoRR*, abs/1908.07245.

Anubha Kabra, Ayush Chopra, Nikaash Puri, Pinkesh Badjatiya, Sukriti Verma, Piyush Gupta, and Balaji Krishnamurthy. 2020. Mixboost: Synthetic oversampling with boosted mixup for handling extreme imbalance. *CoRR*, abs/2009.01571.

Adam Kilgarriff. 2004. How dominant is the commonest sense of a word? In *Text, Speech and Dialogue*, pages 103–111, Berlin, Heidelberg. Springer Berlin Heidelberg.

Jin-Ha Lee, Muhammad Zaigham Zaheer, Marcella Astrid, and Seung-Ik Lee. 2020. Smoothmix: A simple yet effective data augmentation to train robust classifiers. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR) Workshops*.

Jiabin Liu, Bo Wang, Xin Shen, Zhiquan Qi, and Yingjie Tian. 2021. Two-stage training for learning from label proportions. *CoRR*, abs/2105.10635.

Marco Maru, Simone Conia, Michele Bevilacqua, and Roberto Navigli. 2022. Nibbling at the hard core of Word Sense Disambiguation. In *Proceedings of the*

60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 4724–4737, Dublin, Ireland. Association for Computational Linguistics.

George A. Miller, Claudia Leacock, Randee Tengi, and Ross T. Bunker. 1993. A semantic concordance. In *Proceedings of the Workshop on Human Language Technology*, HLT '93, page 303–308, USA. Association for Computational Linguistics.

Andrea Moro and Roberto Navigli. 2015. SemEval-2015 task 13: Multilingual all-words sense disambiguation and entity linking. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 288–297, Denver, Colorado. Association for Computational Linguistics.

Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. SemEval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 222–231, Atlanta, Georgia, USA. Association for Computational Linguistics.

Martha Palmer, Christiane Fellbaum, Scott Cotton, Lauren Delfs, and Hoa Trang Dang. 2001. English tasks: All-words and verb lexical sample. In *Proceedings of SENSEVAL-2 Second International Workshop on Evaluating Word Sense Disambiguation Systems*, pages 21–24, Toulouse, France. Association for Computational Linguistics.

Sameer Pradhan, Edward Loper, Dmitriy Dligach, and Martha Palmer. 2007. SemEval-2007 task-17: English lexical sample, SRL and all words. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 87–92, Prague, Czech Republic. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J. Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *CoRR*, abs/1910.10683.

Benjamin Snyder and Martha Palmer. 2004. The English all-words task. In *Proceedings of SENSEVAL-3, the Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, pages 41–43, Barcelona, Spain. Association for Computational Linguistics.

Kaveh Taghipour and Hwee Tou Ng. 2015. One million sense-tagged instances for word sense disambiguation and induction. In *Proceedings of the Nineteenth Conference on Computational Natural Language Learning*, pages 338–344, Beijing, China. Association for Computational Linguistics.

Laurens Van Der Maaten. 2014. Accelerating t-sne using tree-based algorithms. *J. Mach. Learn. Res.*, 15(1):3221–3245.

Sen Wu, Hongyang R. Zhang, Gregory Valiant, and Christopher Ré. 2020. On the generalization effects of linear transformations in data augmentation. *CoRR*, abs/2005.00695.

Soyoung Yoon, Gyuwan Kim, and Kyumin Park. 2021. Ssmix: Saliency-based span mixup for text classification. *CoRR*, abs/2106.08062.

Hongyi Zhang, Moustapha Cissé, Yann N. Dauphin, and David Lopez-Paz. 2017. mixup: Beyond empirical risk minimization. *CoRR*, abs/1710.09412.

A Appendix

A.1 t-SNE Plots

We present additional t-SNE plots in Figure 4 to show that the augmentation data generated by our method retain the label of the original data. The method of obtaining the t-SNE plots is the same as in section 4.3.

A.2 Augmentation Examples

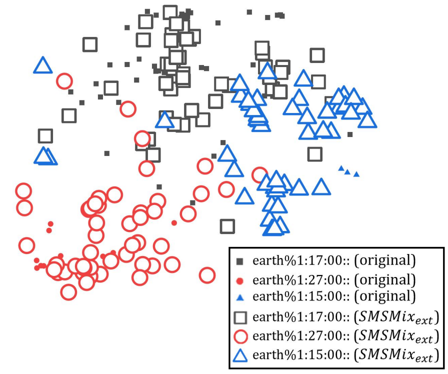
We show examples of augmented sentences for several senses in Tables 3 and 4. Table 3 shows some examples of well-augmented results using our proposed SMSMix. The chosen span with a target word, determined by the saliency score from the original sentence, smoothly blends in with a randomly sampled Wikipedia sentence. On the other hand, the examples shown in Table 4 are cherry-picked failure examples. There are a few failure cases where there are incorrect grammar uses or the sentence does not make sense due to the high discrepancy of context between the two sentences.

A.3 Standard Deviation of Results

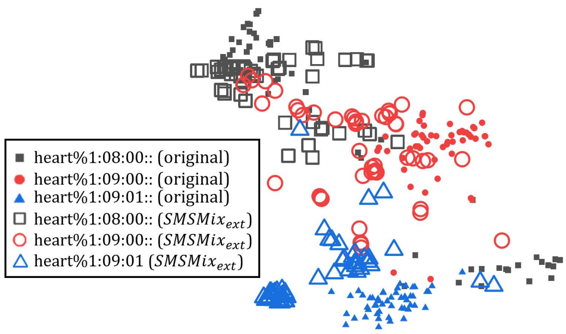
In Table 5 and Table 6, we report the standard deviation of the results obtained in Table 1 and Table 2, respectively. The values were obtained by running the same experiments with five different random seeds.

A.4 Location for Span Injection

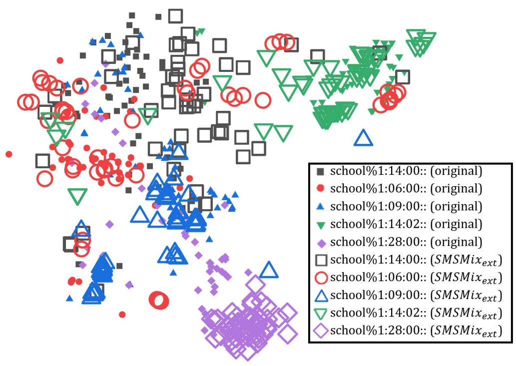
We randomly choose where to inject the sense-maintained span in a sentence, as we use T5's span prediction capability to complete the sentence. To demonstrate the effectiveness of T5's span prediction for SMSMix, we use the example in Figure 2 to create various augmented sentences by injecting the sense-maintained span into different random locations (Figure 5).



(a) earth (noun)



(b) heart (noun)



(c) school (noun)

Figure 4: Latent space visualization showing that SMSMix maintains the sense lablel of the target word.

	Examples
target sense	output%1:10:02:: (signal that comes out of an electronic system)
original	***Outputs*** *of the two systems are measured* by a pulse timing circuit and a resistance bridge, followed by a simple analogue computer which feeds a multichannel recorder.
Wikipedia	The book tells the story of Hendrix and his life through reproductions of rare material such as letters, drawings, postcards and posters.
SMSMix	The book tells the story of how the inputs and ***outputs*** *of the two systems are measured* through rare material such as letters, drawings, postcards and posters.
target sense	work%1:04:01:: (the occupation for which you are paid)
original	In a few places cooperative programs between schools and employers in *clerical* ***work*** *have* shown the same possibilities for allowing the student, while still in school, to develop skills which are immediately marketable upon graduation.
Wikipedia	Throughout his career, he was the recipient of more than 30 awards and honors related to engineering, manufacturing, and the development of heavy equipment.
SMSMix	Throughout his career, his engineering skills and *clerical* ***work*** *have* been the recipient of more than 30 awards and honors related to engineering, manufacturing ,and the development of heavy equipment.
target sense	work%2:38:00:: (proceed along a path)
original	Several photographs and charts of galaxies help the non-scientist keep up with the discussion, and the smooth language indicates the contributors were determined to avoid the *jargon that seems to* ***work*** *its way into* almost every field.
Wikipedia	Barbour gave his time free for the next 25 summers to manage field parties throughout the state, surveying the geological and paleontological resources of the State of Nebraska.
SMSMix	Barbour gave his time free for the next 25 summers to deal with all the *jargon that seems to* ***work*** *its way into* the process of surveying the geological and paleontological resources of the State of Nebraska.
target sense	condition%1:10:01:: (an assumption on which rests the validity or effect of something else)
original	This is what we mean when we say this demand must be accepted *without* ***condition.***
Wikipedia	The Society was wound up in the year 2001 when no ordinary members wanted to be nominated as new committee members.
SMSMix	The Society is open to all, *without* ***condition*** , except in the year 2001 when no ordinary members wanted to be nominated as new committee members.
target sense	lighting%1:06:00:: (apparatus for supplying artificial light effects for the stage or a film)
original	When improvements are recommended in *working conditions - such as* ***lighting*** , *rest rooms* , eating facilities , air-conditioning - do you try to set a measure of their effectiveness on productivity?
Wikipedia	However, when de Gaulle first introduced the Fouchet Plan in 1961, it faced opposition from many of the member states.
SMSMix	However, when de Gaulle first introduced the idea, improvements are recommended in *working conditions - such as* ***lighting****, rest rooms* and transport between member states.

Table 3: Examples of well-augmented sentences, generated by SMSMix. Red represents the sense-maintained span in the original sentence, blue represents a random span in an external sentence that is going to be replaced, and green represents the span prediction made by T5.

	Examples
target sense	make%2:40:02:: (achieve a point or goal)
original	It is interesting to note that medium *compulsives in the unstructured schools* ***made*** *lowest achievement scores* (although not significantly lower) .
Wikipedia	Josh tries to distance himself as much as possible from her, for fear of what might happen if she finds out what he is.
SMSMix	Josh tries to distance himself as much as possible from her, for fear of losing her, but the *compulsives in the unstructured schools* ***made*** *the lowest achievement scores* in the.
target sense	employment%1:04:01:: (the act of using)
original	Another case may be given in illustration of a successful use *of analysis, and also of the* ***employment*** *of a procedure* for intensive analysis.
Wikipedia	Ahh!, which was released on May 21, 1988 and would ultimately go on to sell 8 million copies worldwide.
SMSMix	Ahh! *of analysis, and also of the* ***employment*** *of a procedure* that would ultimately go on to sell 8 million copies worldwide.
target sense	replace%2:41:00:: (take the place or move into the position of)
original	This and raw *sugar* ***replace*** *ordinary* refined sugar on the tables and very little sugar is used in cooking.
Wikipedia	These are public housing units and estates aimed at Singaporeans who do not want a HDB flat but might find private property too expensive.
SMSMix	These are public housing units and estates aimed at Singaporeans who want to *sugar* ***replace*** *ordinary* sugar a HDB flat but might find private property too expensive.
target sense	people%1:14:03:: (the common people generally)
original	Linguists have not always been more enlightened than “*practical* ***people***” *and sometimes have insisted on incredibly trivial* points while neglecting things of much greater significance.
Wikipedia	According to the law of marginal utility, the value of each good in a stock of identical goods is utility of the last and most easily dispensable unit.
SMSMix	According to the law of marginal utility , the value of each good in a stock of identical goods is “the most *practical* ***people***” *and sometimes have insisted on incredibly trivial* details.
target sense	shift%2:38:02:: (move around)
original	Important as these differences are, they should not obscure the basic fact that *by* ***shifting*** *the hypothalamic balance* sufficiently to the parasympathetic side, we produce depressions, whereas a shift in the opposite direction causes excitatory effects and, eventually, maniclike changes.
Wikipedia	The district is the mining and forestry centre of Suriname, with many large bauxite mining operations operating.
SMSMix	The district is the mining and forestry centre of Suriname, with many mines *by* ***shifting*** *the hypothalamic balance* of.

Table 4: Failure examples of augmentation by SMSMix. Red represents the sense-maintained span in the original sentence, blue represents a random span in an external sentence that is going to be replaced, and green represents the span prediction made by T5.

Target Sense giveaway%1:21:00:: (a gift of public land or resources for the private gain of a limited group)

Original Sentence Have you permitted it to become a giveaway program rather than one that has the goal of improved employee morale and , consequently , increased productivity?

External Sentence In 2005, Disown secured a spot on the OzzFest 2004 tour performing in front of more than 30,000 people.

SMSMix In 2005, Disown secured a spot on a giveaway program for a 2004 tour performing in front of more than 30,000 people.

External Sentence In 2005, Disown secured a spot on the OzzFest 2004 tour performing in front of more than 30,000 people.

SMSMix In 2005, Disown secured a spot on the Billboard Music Awards through a giveaway program, performing in front of more than 30,000 people.

External Sentence In 2005, Disown secured a spot on the OzzFest 2004 tour performing in front of more than 30,000 people.

SMSMix In 2005, Disown secured a spot on the OzzFest 2004 tour as part of a giveaway program with an audience of more than 30,000 people.

External Sentence In 2005, Disown secured a spot on the OzzFest 2004 tour performing in front of more than 30,000 people.

SMSMix In 2005, Disown secured a spot on the OzzFest 2004 tour performing in front of thousands of people and a giveaway program to select people.

Figure 5: Examples of SMSMix using various different injection locations. Red represents the sense-maintained span in the original sentence, blue represents a random span in an external sentence that is going to be replaced, and green represents the span prediction made by T5.

Methods	Dev Set		Test Sets									
	SE07		SE2		SE3		SE13		SE15		ALL	
	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1
+Oversample	0.38	0.41	0.10	0.10	0.11	0.18	0.10	0.20	0.06	0.10	0.49	0.10
+SMSMix$_{int}$	0.15	0.18	0.09	0.11	0.18	0.20	0.23	0.30	0.10	0.06	0.05	0.05
+SMSMix$_{ext}$	0.19	0.22	0.04	0.09	0.08	0.06	0.09	0.15	0.08	0.09	0.05	0.05

Table 5: Standard deviation values for the experimental results in Table 1. The values were obtained by running the same experiments with five different random seeds.

	MFS		LFS		0-lex		0-lex-def		0-def	
	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1	m-F1	M-F1
+ Oversample	0.11	0.17	0.14	0.15	0.12	0.15	0.15	0.13	0.18	0.25
+ SMSMix$_{int}$	0.07	0.06	0.19	0.10	0.04	0.10	0.10	0.13	0.10	0.13
+ SMSMix$_{ext}$	0.08	0.04	0.12	0.12	0.04	0.08	0.10	0.07	0.08	0.07

Table 6: Standard deviation values for the experimental results in Table 2. The values were obtained by running the same experiments with five different random seeds.