Chinese Content Scoring: Open-Access Data Sets and Features on Different Segmentation Levels

Yuning Ding **Andrea Horbach** **Haoshi Wang** **Xuefeng Song** **Torsten Zesch**
Language Technology Lab, University Duisburg-Essen
(yuning.ding|andrea.horbach|torsten.zesch)@uni-due.de

Abstract

In this paper, we analyse the challenges of Chinese content scoring in comparison to English. As a review of prior work for Chinese content scoring shows a lack of open-access data in the field, we present two short-answer data sets for Chinese. The Chinese Educational Short Answers data set (CESA) contains 1800 student answers for five science-related questions. As a second data set, we collected ASAP-ZH with 942 answers by re-using three existing prompts from the ASAP data set.

We adapt a state-of-the-art content scoring system for Chinese and evaluate it in several settings on these data sets. Results show that features on lower segmentation levels such as character n-grams tend to have better performance than features on token level.

1 Introduction

Short answer questions are a type of educational assessment that requires respondents to give natural language answers in response to a question or some reading material (Rademakers et al., 2005). The applications used to automatically score such questions are usually thought of as content scoring systems, because content (and not linguistic form) is taken into consideration for automatic scoring (Ziai et al., 2012). While there is a large research body for English content scoring, there is less research for Chinese.[1] The largest obstacle for more research on Chinese is the lack of publicly available data sets of Chinese short answer questions.

Working with Chinese poses substantially different challenges than work on English data. Unlike English, which uses spaces as natural separators between words, segmentation of Chinese texts into tokens is challenging (Chen and Liu, 1992). Furthermore, there are more options on which level to segment Chinese text. Apart from tokenization and segmentation into characters, which are two options also available and often used for English, segmentation into components, radicals and even individual strokes are additionally possible for Chinese. Table 1 gives an example for the segmentation options in both languages. Orthographic variance can be challenging in both languages, but behaves very differently. Non-word errors, which is the main source of orthographic problems in English (Mitton, 1987), can by definition not happen in Chinese, due to the input modalities.

Language	Level	Unigrams
English	word	panda
	characters	p, a, n, d, a
Chinese	word	熊猫
	characters	熊，猫
	components	能，灬，犭，苗
	radicals	灬，犭
	strokes	㇜ ㇔ ㇑ ㇆ ㇐ ㇐ ...

Table 1: Comparison of segmentation possibilities in English and Chinese

In the remainder of this paper, we will discuss these challenges in more detail (Section 2). We review prior work on Chinese content scoring (Section 3) and present two new freely-available data sets of short answers in Chinese (Section 4). In Section 5, we adapt a

[1] In this work, we use the term ‘Chinese’ as abbreviation for Mandarin Chinese, which includes simplified and traditional written Chinese. Cantonese, Wu, Min Nan and other dialects are not included.

Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing, pages 347–357
December 4 - 7, 2020. ©2020 Association for Computational Linguistics

machine learning pipeline for automatic scoring with state-of-art NLP tools for Chinese. We investigate the extraction of n-gram features on all possible segmentation levels. In addition, we use features based on the Pinyin transcription of Chinese texts and experiment with the removal of auxiliary words as an equivalent to lemmatization in English. We evaluate these features on our new data sets as well as, for comparison, an English data set translated into Chinese.

2 Challenges in Chinese Content Scoring

In this section, we highlight the main challenges when processing Chinese learner data in comparison to English data sets. We first focus on segmentation, as tokenization is more difficult in Chinese than in English and there are more linguistic levels on which to segment a Chinese text compared to English. Next, we discuss variance in learner answers, which is a challenge for content scoring in any language but manifests itself in Chinese differently than in English.

2.1 Segmentation

English has an alphabetic writing system with some degree of grapheme-to-phoneme correspondence. The Chinese language, in contrast, uses a logosyllabic writing system, where characters represent lexical morphemes. Chinese words can be formed by one or more characters (Chen, 1992). Unlike English, where words are separated by white-spaces, the fact that Chinese writing does not mark word boundaries makes word segmentation a much harder task in Chinese NLP (e.g., Chen and Liu (1992); Huang et al. (1996)). According to a recent literature review on Chinese word segmentation (Zhao et al., 2019), the best-performing segmentation tool has an average F1-value of only around 97%. A major challenge is the handling of out-of-vocabulary words.

In English content scoring, word level features such as word n-grams or word embeddings have proven to be effective (e.g., Sakaguchi et al. (2015); Riordan et al. (2017)). Additionally, character features are frequently used to capture orthographic as well as morphological variance (e.g., Heilman and Madnani (2013); Zesch et al. (2015)).

In the light of the tokenziation challenges mentioned above, it is surprising that although most prior work on Chinese also applies word-level features (see Section 3), the performance of their tokenizers are barely discussed and character-level features are neglected altogether.

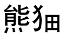
Apart from words and characters, there are more possibilities of segmentation in Chinese as discussed above. Consider, for example, a Chinese bi-morphemic word such as $\overset{\text{panda bear}}{\text{熊猫}}$. It can additionally be segmented on the stroke, component and radical level as shown in Table 1.

It has been argued that the morphological information of characters in Chinese consists of the sequential information hidden in stroke order and the spatial information hidden in character components (Tao et al., 2019). Each Chinese character can directly be mapped into a series of strokes (with a particular order). On the component level, it has been estimated that about 80% of modern Chinese characters are phonetic-logographic compounds, each of which consists of two components: One carries the sound of the character (the stem) and the other the meaning of the character (the radical) (Li, 1977). We argue that, together with strokes, both kinds of components may be used as features in content scoring. Note that in some cases, a character has only one component, which in the extreme case consists of one stroke only, so that for the character $\overset{\text{one}}{\text{一}}$, all four segmentation levels yield the same result, somewhat comparable to an English one-character word, such as "I".

2.2 Linguistic Variance

Variance in learner answers has a major influence on content scoring performance (Horbach and Zesch, 2019), i.e., the more variance between the answers to a specific prompt, the harder it is to score automatically. If we ignore cases of conceptually different answers, variance means different realizations with approximately the same semantic meaning. As shown in Table 2, if we have a question about the eating habits of pandas, Chinese short answers can contain similar variance as in English, which is realized as both orthographic variance caused by spelling errors as well as variance of linguistic expression. Note that these types of variance should not influence the score of an answer as it depends only from the content of the answer. Both types of variance are further discussed in the following.

Spelling errors in English can be classified into non-word and real-word spelling errors. In our example, "bambu" is a non-word, while "beer" is a real word spelling error. Both error types occur frequently in English short answer data sets, with non-word errors being more frequent (Mitton, 1987, 1996). A content scoring system must therefore be able to generalize by taking variance in spelling into account (Leacock and Chodorow, 2003). To do so, many systems for English data use character-level features (Heilman and Madnani, 2013; Horbach et al., 2017), such that "bamboo" and "bambu", while being different tokens, share, for example, the character 3-grams ‘bam’ and ‘amb’ .

For Chinese, the situation is entirely different. Non-word spelling errors are rare and even impossible for digitized data because of the input modalities typically used for Chinese text. When entering a Chinese text on the computer, a writer would normally type the phonetic transcription Pinyin, which is the Romanization of Chinese characters based on their pronunciation. After typing a Pinyin, the writer is shown all corresponding characters from which they choose the right one. As this selection list contains only valid Chinese characters, non-word errors cannot occur by definition. Even if the original data set was collected in hand-written format, the transcription process forces transcribers to correct any non-word error that might occur in the data. For example, if the learner accidentally wrote panda bear 熊猫 as , the transcriber has no choice but to correct such an error, since the non-word character simply does not exist in the Chinese character set.

There are two steps in the writing / transcription process where errors can still occur: typing letters to spell a Pinyin and choosing a character out of a list for this Pinyin. Previous experiments showed that people usually do not check Pinyin for errors, but wait until the Chinese characters start to show up (Chen and Lee, 2000). This behaviour generates two types of real-word spelling errors. In our example, spelling errors like confusing poor 穷 (qiǒng) with bear 熊 (xiǒng) are normally caused by wrong letters typed in the first step. The other error type, i.e., choosing a wrong word from the homophones, leads to spelling errors like pearl 珠子(zhū zi) instead of bamboo 竹子 (zhú zi). Researchers found that nearly 95% of errors are due to the misuse of homophones (Yang et al., 2012), i.e., are errors of the second type. In order to reduce the influence of these errors in content scoring, introducing features presented as Pinyin might be beneficial.

Variance of linguistic expression is obviously found in both English and Chinese short answers. As shown in Table 2, nearly the same content can be expressed using different lexical and syntactic choices. Human annotators can usually abstract away from these differences and treat all answers the same. However, linguistic variance is a challenge for automatic scoring systems.

In English content scoring, lemmatization is often considered a useful method to reduce part of the variance (Koleva et al., 2014). In this process, words are reduced to their base forms, such as substituting "ate" with "eat" and deleting the "s" after "bamboo". In Chinese, similar grammatical morphemes such as "了" and "们", termed auxiliary words (Zan and Zhu, 2009), which indicate the past tense and plural, can also be deleted in a preprocessing step to achieve a similar effect.

Another type of variance is caused by synonyms. For such cases of lexical variance, external knowledge is often needed to decide that two different words are interchangeable. However, as we can see in Table 2, some synonyms, such as "panda bears" vs. "pandas" and bamboo 竹子 vs. bamboo 竹 share some character(s). Such similarities can be covered by character features, but not token n-grams.

In summary, there is the challenge of the segmentation of Chinese texts into tokens. Features extracted on other segmentation levels might be more robust and therefore helpful for automatic scoring. At the same time, NLP techniques which are useful to reduce variance

	English	Chinese
Reference Answer	Panda bears eat bamboo.	panda bear eat bamboo 熊猫 吃 竹子 。
Orthographic Variance	Panda beers eat bambu.	poor cat eat pearl 穷 猫 吃 珠子。
Expression Variance	Panda bears ate bamboos.	panda bear eat <grammatical morpheme for past tense> 熊 猫 吃 过 bamboo <grammatical morpheme for plural> 竹子 们 。
	Pandas eat bamboo.	panda bear eat bamboo 熊猫 吃 竹 。

Table 2: Example answers showing variance in English and Chinese for the question: *What do panda bears eat?*

in English, especially lemmatization, have not yet been transferred to Chinese. Thus, we will explore in our experiments both n-gram features on different levels and the removal of auxiliary words.

3 Prior Work on Chinese Content Scoring

As shown in Table 3, all prior work on Chinese content scoring uses lexical features on the word level, such as word n-grams and sentence length in tokens. They are not only used in shallow learning methods like support vector machines (SVM) or support vector regression (SVR) (Wang et al., 2008; Wu and Shih, 2018), but also applied to deep learning methods like long-short term memory recurrent neural networks (LSTM) (Yang et al., 2017; Huang et al., 2018) or deep autoencoders (Yang et al., 2018).

Also for neural models using word embeddings, word-level tokenization is necessary. Wu and Yeh (2019) train 300-dimensional word2vec word embeddings on sentences from their data set along with Chinese Wikipedia articles and classify student answers with a convolution neural network (CNN). Li et al. (2019) use a Bidirectional Long Short-Term Memory (Bi-LSTM) network for semantic feature extraction from pre-trained 300-dimensional word embeddings (Li et al., 2018) and score student answers based on their similarity to the reference answer using a mutual attention mechanism.

For segmentation, most prior work uses the jieba tokenizer [2] for pre-processing. However, the performance of the tokenization is rarely discussed. We also notice that no related work uses segmentation on character or component level. Yang et al. (2018) perform stop word removal, but they do not mention if it included some kind of removal of grammatical markers.

4 Chinese Scoring Data Sets

In this section, we review existing Chinese content scoring data sets. They are not publicly available, which is a major obstacle to reproducibility in the field. We thus produce two new Chinese data sets (see detailed description in Section 4.2), which are available online[3] to foster future research .

4.1 Existing Data Sets

Horbach and Zesch (2019) give an overview of publicly available data sets for content scoring, five of which are for English, and compare them based on properties such as prompt type, learner population and data set size.

Unfortunately, we did not find any freely available Chinese content scoring data sets. Since we could not access the data sets used in related work, we can only compare them based on their brief descriptions, according to the aspects of comparison mentioned above. Results are shown in Table 4.

The Debris Flow Hazard (DFH) data set is used in the earliest work. It contains more than 1000 answers for 2 prompts in a creative problem-solving task. The learner population are high-school students from Taiwan, who speak native Chinese (Wang et al., 2008).

[2] https://github.com/fxsjy/jieba

[3] https://github.com/ltl-ude/ChineseShortAnswerDatasets

Reference	Data Set	Preprocessing	Features	Classifier	Evaluation
Wang et al. (2008)	DFH task	tokenization, POS tagging	word uni-/bigrams, POS bigrams	SVM	r=.92
Wu and Shih (2018)	SCB-ZH^{MT} CS-EN^{MT}	tokenization (jieba)	sentence length, word unigrams, BLEU score	SVR, SVM	acc=.60 RMSE=1.17
Yang et al. (2017)	CRCC	tokenization (jieba)	word unigrams	LSTM	acc=.76, Cohen's κ=.61
Yang et al. (2018)	CRCC	punctuation and stop word removal, tokenization (jieba)	word unigrams	Auto-encoder	acc=.74, qwk=.64
Huang et al. (2018)	CRCC	tokenization (jieba)	word vector trained on CBOW	LSTM	acc=.74, qwk=.62
Wu and Yeh (2019)	ML_SQA SCB-ZH^{MT}	tokenization (jieba)	300D pre-trained word embedding	CNN	acc=.91, recall=.82
Li et al. (2019)	Law Questions	tokenization	300D pre-trained word embedding	Bi-LSTM	acc=.88

Table 3: Overview of related work in Chinese content scoring.

The Chinese Reading Comprehension Corpus (CRCC) (Yang et al., 2018), contains five reading comprehension questions. Each question has on average 2500 answers from students in grade 8.

Instead of collecting and annotating a data set from scratch, Wu and Shih (2018) translated the English SciEntBank (Dzikovska et al., 2013) and the computer science (CS) (Mohler and Mihalcea, 2009) data sets to Chinese. The data set was first translated using machine translation. In order to solve word usage and grammar problems, 12% of the sentences were manually corrected. In their most recent work, the authors also collected a data set with 12 short answer questions and overall 600 answers related to machine learning (ML_SQA) to compare with the CS-ZH^{MT} data set (Wu and Yeh, 2019).

In the most recent work (Li et al., 2019), a large data set containing 85.000 student and reference answers was collected from a national specialty examination related to law.

4.2 Collection of Open-access Data Sets

As part of the contribution in this paper, we collected two new data sets for Chinese content scoring: Chinese Short Answer (CESA) and ASAP-ZH. In addition, we provide a machine-translated version of the the original ASAP-SAS English data, ASAP-ZH^{MT}. Table 4 shows key properties, while Table 5 gives example answers of each data set.

Chinese Educational Short Answers (CESA) contains five questions from the physics and computer science domain (see Table 6). Answers are collected from 360 students in the computer science department of Zhengzhou University. Each participant was required to answer each question with a maximum of 20 characters, resulting in an average answer length of 13.5 characters. Two annotators speaking native Chinese with computer science background scored the answers into three classes, 0, 1 and 2 points, with an average inter-annotator agreement of 0.9 quadratically weighted kappa (QWK).

ASAP-ZH This data set is based on the ASAP short-answer scoring data set released by the Hewlett Foundation.[4] ASAP contains ten short answer prompts covering different subjects and about 2000 student answers per prompt. Prompt 1, 2 and 10 are science-related tasks, which do not have a strong cultural background, and are therefore considered as appropriate to be transferred to other languages.

Therefore, we collected answers in Chinese

[4] http://www.kaggle.com/c/asap-sas

Data Set	Type	#Answers	#Prompts	Labels	Level
DFH	creative problem solving	2,698	2	[0,1,...,28]	high school
CRCC	reading comprehension	12,528	5	[0,1,2,3,(4),(5)]	middle school
SciEntsBank-ZH^{MT}	science	9,804	197	binary&diagnostic	high school
CS-ZH^{MT}	computer science	630	21	[0, 0.5,..., 5]	university
ML_SQA	computer science	608	12	binary	university
Law Questions	law	85,000	2	[0,1.5,3]/[0,1,1.5]	-
CESA	physics, computer science	1,800	5	[0,1,2]	university
ASAP-ZH	science	942	3	[0,1,2,(3)]	high school
ASAP-ZH^{MT}	science	6,119	3	[0,1,2,(3)]	high school

Table 4: Chinese content scoring data sets: data sets from previous work (upper part) and our new data sets (lower part)

for these three prompts after manually translating the prompt material. The data collection provider BasicFinder[5] helped us to collect 942 answers altogether, 314 answers for each prompt. They are collected from students in high school from grades 9-12, which is comparable with the set of English answers in the ASAP-SAS data set. The answers are transcribed into digital form manually after being collected in handwriting. After reaching an acceptable agreement on a set of answers from the original ASAP-SAS, two annotators speaking native Chinese scored the ASAP-ZH data on a scale from 0 to 3 points (prompt 1 and 2) or 0 to 2 points (prompt 10) with an average QWK of 0.7. Key statistics for the data set can be found in Table 7.

ASAP-ZH^{MT} For comparison, we also translated the English answers in prompts 1,2 and 10 in the original ASAP-SAS data set to Chinese using the Google Translate API.[6] The examples in Table 5 show that some translation errors can be found, especially when errors exist already in the original text. Words containing spelling errors like “wat” instead of ”what” are simply not translated at all. The overall translation quality is also not perfect, for example, the word “coolest” is wrongly translated into $\overset{\text{most}}{\text{最}}$ $\overset{\text{fashioned}}{\text{酷的}}$ instead of the correct $\overset{\text{most}}{\text{最}}$ $\overset{\text{coldest}}{\text{冷的}}$.

As shown in Tables 7 and 8, the average length of the translated answers is larger than the length of the original Chinese answers to the same prompt in our re-collected data set. One explanation could be that paid crowd workers are less motivated than actual students and therefore write shorter answers.

5 Experimental Setup

In this section, we adapt a state-of-the-art content scoring system to Chinese. We evaluate it in six settings with different feature sets on the data sets described above in order to investigate different options for segmentation of Chinese text. Table 9 gives an example for the different segmentation options, which will also be detailed in Section 5.2. Additionally, we add a pre-processing step to remove all auxiliary words in the data in order to simulate the effect of lemmatization in English content scoring.

5.1 General Experimental Setup

For all our experiments, we use the ESCRITO (Zesch and Horbach, 2018) toolkit and extended it with readers and tokenization for Chinese text. ESCRITO is a publicly available general-purpose scoring framework based on DKPro TC (Daxenberger et al., 2014), which uses an SVM classifier (Cortes and Vapnik, 1995) using the SMO algorithm as provided by WEKA (Witten et al., 1999). For all kinds of features, we use the top 10000 most frequent

[5] https://www.basicfinder.com/en

[6] https://cloud.google.com/translate

Data Set	ID	Score	Example
CESA	5	2	The machine summarizes a large amount of data and finds the pattern from it 机器总结大量数据，从中找到规律
		1	Machines can learn things by themselves 机器能自己学习东西
		0	Let the machine learn human thinking ability 让机器学习人的思想能力
ASAP-ZH	10	2	White: make the indoor temperature not too high, 白色使室内气温不太高 experiments show that white has the lowest light energy absorption rate 实验表明白色对光的能量吸收率最低
		1	Black allows the doghouse to absorb more heat in the light, making it warm 黑色能让狗窝在光下吸更多的热，使其温暖
		0	Dark gray: keep the temperature unchanged, 深灰色：： 使温度不变， the lighter the color, the lower the temperature 颜色越浅温度越低
ASAP-ZH^{MT}	10	2	white : : having white paint would make the dog house colder, 白色:: 有白色油漆会使狗屋更冷， so in the summer the dog would not be hot. 所以在夏天狗不会很热。 The average for white is the coolest temperature (42 (DEG)) 白色的平均值是最酷的 温度（42（DEG））
		1	black :: Because, the darker the lid color, 黑色:: 因为，盖子颜色越深， the greater the increase in the air temperature in the glass jar. 玻璃罐中空气温度的升高就越大。
		0	light gray :: The light grey will effect the doghouse by making it more noticable 浅灰色：浅灰色会使狗狗更加显眼， and plus dogs can only see black, white and grey. 加上狗只能看到黑色，白色和灰色。

Table 5: Example answers in our data sets.

1- to 5-grams. Due to the limited amount of data, we use 10-fold cross-validation on both data sets.

For evaluation, we use accuracy, i.e., the percentage of student answers scored correctly, as well as QWK, which does not only consider whether an answer is classified correctly or not, but also how far it is from the gold standard classification.

5.2 Feature Sets

Token Baseline As a baseline, we follow previous work and use tokenization as segmentation, based on the HanLP tokenizer (He, 2020).

Pinyin Features In order to reduce the variance caused by spelling errors, we transcribe the text into Pinyin using *cnchar* (Chen, 2020) and extract ngrams on the level of transcribed characters. Note that we did not include information about tones in Pinyin on purpose, in order to cover spelling errors caused by homophones.

Character Features For this segmentation level, we simply split a text into individual characters.

Component Features To extract these features on sub-character level, we use a dictionary with 17,803 Chinese characters[7] and their components to decompose all characters.

Radical Features Remember that radicals are only those components carrying the meaning of characters and might therefore be particularly useful in content scoring. We use XMNLP (Li, 2019) to extract the radicals of each character and use only those as features. Note that some radicals as defined by the "Table of Indexing Chinese Character Compo-

[7] https://github.com/kfcd/chaizi

ID	Prompt	IAA	avg.	Distribution		
		(QWK)	Length	score 0	score 1	score 2
1	why we can use diamond cut glass 为什么 我们 能 用 钻石 切 玻璃?	.94	9.6	60%	38%	2%
2	why red clothes looks as red 为什么 红色 衣服 看起来 是 红色的?	.83	14.7	63%	20%	17%
3	what is artificial intelligence 什么 是 人工 智能 ?	.91	15.3	39%	37%	24%
4	what is natural language 什么 是 自然 语言 ?	.93	12.1	66%	15%	19%
5	what is machine learning 什么 是 机器 学习 ?	.89	15.7	31%	36%	33%

Table 6: Overview of prompts in CESA

ID	IAA	avg.	Distribution			
	(QWK)	Length	score 0	score 1	score 2	score 3
1	.72	35.3	20%	28%	34%	18%
2	.70	38.2	36%	43%	18%	3%
10	.69	37.6	54%	32%	14%	

Table 7: Overview of prompts in ASAP-ZH

ID	IAA	avg.	Distribution			
	(QWK)	Length	score 0	score 1	score 2	score 3
1	.96	68	22%	27%	31%	20%
2	.94	94	14%	25%	36%	25%
10	.91	61	17%	47%	36%	

Table 8: Overview of prompts in ASAP-ZH^{MT}

nents"[8] can consist of more than one component, therefore the radicals are not a proper subset of the components extracted above.

Stroke Features We use the *cnchar* tool to represent each answer as a sequence of individual strokes, following the stroke order for each character. Although we show the strokes in their original shapes in Table 9, a letter encoding is used in the experiment for an efficient processing.

Auxiliary Words Removal Based on the knowledge database released by Han et al. (2011), which contains 45 common auxiliary words in modern Chinese, we remove all these grammatical morphemes on token level to reduce the influence of expression variance. In our example shown in Table 9, the possessive

Answer	diamond ’s hardness great 钻石 的 硬度 大
Tokens	钻石，的，硬度，大
Pinyin	Zuan, Shi, De, Ying, Du, Da
Characters	钻，石，的，硬，度，大
Components	金占，一丿口，白勺， 石更，广廿又，人一
Radicals	钅，石，白，石，广，大
Strokes	㇒㇐㇐㇐㇙㇑㇐㇑㇕㇐, ㇐㇒㇑㇕㇐, ㇒㇑㇕㇐㇐㇒㇆㇔, ㇐㇒㇑㇕㇐㇐㇑㇕㇐㇐㇒㇏, ㇔㇐㇒㇐㇑㇑㇐㇇㇑㇖㇏, ㇐㇒㇏

Table 9: Different segmentation levels for an answer in CESA, prompt 1.

marker ‘s 的 is eliminated.

6 Results and Discussion

Table 10 shows the performance of the different system configurations for the individual data sets, per prompt as well as averaged over all prompts from the same data set. First, we see that all feature sets were able to learn something meaningful from the training data. Although the performance of different feature sets is quite close to each other, we see a slight but significant advantage across data sets of component and character features over the token baseline.

In order to check if tokenization caused

[8] http://www.moe.gov.cn/s78/A19/yxs_left/moe_810/s230/201001/t20100115_75694.html

Data Set	CESA						ASAP-ZH				ASAP-ZH^{MT}			
Prompt	1	2	3	4	5	**avg.**	1	2	10	**avg.**	1	2	10	**avg.**
Token	.91	.84	.59	.66	.48	.70	.54	.40	.50	.48	.66	.59	.63	.63
Pinyin	-.02	+.03	-.03	+.01	-.03	+.01**	+.13	+.01	+.04	+.09**	-.02	+.01	+.01	±0
Character	-.01	+.03	±0	+.11	+.05	**+.04****	+.13	+.03	+.06	+.07**	±0	+.04	+.04	**+.02***
Component	-.03	+.03	-.01	+.10	+.02	+.02**	+.17	+.04	+.08	**+.10****	-.01	±0	+.04	+.01**
Radical	-.02	+.02	+.03	+.07	±0	+.02**	+.08	+.08	+.02	+.06**	+.02	-.02	+.04	+.01
Stroke	-.01	-.02	-.02	+.06	-.04	-.01**	+.14	+.07	+.04	+.08**	-.01	-.02	-.03	-.02**
- Auxiliary	±0	±0	+.03	+.02	+.01	+.01**	-.01	±0	-.01	-.01**	-.01	-.01	+.01	-.01**

$^{**}p < 0.01$, $^{*}p < 0.05$

Table 10: Classification results on different feature sets in QWK values.

problems in scoring, we manually inspected 100 answers from prompt 1 and 4 in CESA. However, we found that tokenization was only erroneous in 12 cases. Surprisingly, most of them occurred in prompt 1, where the token baseline even outperformed the character features and not in prompt 4, where character features performed better.

We also had a closer look at a number of student answers which are assigned a wrong score by the token baseline model but not by models with more fine-grained features. 7 out of 18 instances contain indeed variants of more frequent words in the data set. For example, $\overset{\text{human}}{\text{人们}}$ and $\overset{\text{human}}{\text{人}}$ are less-frequently seen variants of $\overset{\text{human}}{\text{人类}}$, all of which are indicators of a correct answer. This supports the assumption that, like in English, character-level features can capture variance in learner answers, in this case by handling variance in lexical choice.

The usage of Pinyin did not bring the expected benefit, possibly because the amount of spelling errors is not substantial enough in the data. Similarly, removing auxiliary words appears to have little influence on scoring performance.

7 Summary & Future Work

In this paper, we discussed the main challenges in Chinese content scoring in comparison with English, namely segmentation and a different form of linguistic variance. We reviewed related work in Chinese content scoring and saw a need for open-access scoring data sets in Chinese. Therefore, we collected two new data sets, CESA and ASAP-ZH, and release them for research in the future.

While previous work has been limited to word-level features, we conducted a comparison of features on different segmentation levels. Although the difference between feature sets was in general small, we found that some answers with unusual expressions have a tendency to be better scored with models trained on lower level features, such as character n-grams.

In the future, we will extend our comparison of segmentation levels also to a deep learning setting, using embeddings of different granularity (Yin et al., 2016).

References

Hsuan-Chih Chen. 1992. Reading comprehension in chinese: Implications from character reading times. *Language processing in Chinese*, pages 175–205.

Keh-Jiann Chen and Shing-Huan Liu. 1992. Word identification for mandarin chinese sentences. In *Proceedings of the 14th conference on Computational linguistics-Volume 1*, pages 101–107. Association for Computational Linguistics.

Tack Chen. 2020. Full-featured, multi-end support for hanyu pinyin strokes js library.

Zheng Chen and Kai-Fu Lee. 2000. A new statistical approach to chinese pinyin input. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 241–247.

Corinna Cortes and Vladimir Vapnik. 1995. Support-vector networks. *Machine learning*, 20(3):273–297.

Johannes Daxenberger, Oliver Ferschke, Iryna Gurevych, and Torsten Zesch. 2014. Dkpro tc: A java-based framework for supervised learning experiments on textual data. In *Proceedings of 52nd Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 61–66.

Myroslava O Dzikovska, Rodney D Nielsen, Chris Brew, Claudia Leacock, Danilo Giampiccolo, Luisa Bentivogli, Peter Clark, Ido Dagan, and Hoa T Dang. 2013. Semeval-2013 task 7: The joint student response analysis and 8th recognizing textual entailment challenge. Technical report, NORTH TEXAS STATE UNIV DENTON.

Ying-Jie Han, Hong-Ying Zan, Kun-Li Zhang, and Yu-Mei Chai. 2011. Automatic annotation of auxiliary words usage in rule-based chinese language. *Jisuanji Yingyong/ Journal of Computer Applications*, 31(12):3271–3274.

Han He. 2020. HanLP: Han Language Processing.

Michael Heilman and Nitin Madnani. 2013. Ets: Domain adaptation and stacking for short answer scoring. In *Second Joint Conference on Lexical and Computational Semantics (* SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, pages 275–279.

Andrea Horbach, Yuning Ding, and Torsten Zesch. 2017. The influence of spelling errors on content scoring performance. In *Proceedings of the 4th Workshop on Natural Language Processing Techniques for Educational Applications (NLPTEA 2017)*, pages 45–53.

Andrea Horbach and Torsten Zesch. 2019. The influence of variance in learner answers on automatic content scoring. *Frontiers in Education*, 4:28.

Chu-Ren Huang, Keh-Jiann Chen, and Li-Li Chang. 1996. Segmentation standard for chinese natural language processing. In *Proceedings of the 16th Conference on Computational Linguistics - Volume 2*, COLING ’ 96, page 1045–1048, USA. Association for Computational Linguistics.

Yuwei Huang, Xi Yang, Fuzhen Zhuang, Lishan Zhang, and Shengquan Yu. 2018. Automatic chinese reading comprehension grading by lstm with knowledge adaptation. In *Pacific-Asia Conference on Knowledge Discovery and Data Mining*, pages 118–129. Springer.

Nikolina Koleva, Andrea Horbach, Alexis Palmer, Simon Ostermann, and Manfred Pinkal. 2014. Paraphrase detection for short answer scoring. In *Proceedings of the third workshop on NLP for computer-assisted language learning*, pages 59–73.

Claudia Leacock and Martin Chodorow. 2003. C-rater: Automated scoring of short-answer questions. *Computers and the Humanities*, 37(4):389–405.

Dongjin Li, Tianyuan Liu, Wei Pan, Xiaoyue Liu, Yuqing Sun, and Feng Yuan. 2019. Grading chinese answers on specialty subjective questions. In *CCF Conference on Computer Supported Cooperative Work and Social Computing*, pages 670–682. Springer.

HT Li. 1977. The history of chinese characters. *Taipei, Taiwan: Lian-Jian.*

Shen Li, Zhe Zhao, Renfen Hu, Wensi Li, Tao Liu, and Xiaoyong Du. 2018. Analogical reasoning on chinese morphological and semantic relations. *arXiv preprint arXiv:1805.06504*.

Xianming Li. 2019. A lightweight chinese natural language processing toolkit. `https://github.com/SeanLee97/xmnlp`.

Roger Mitton. 1987. Spelling checkers, spelling correctors and the misspellings of poor spellers. *Information processing & management*, 23(5):495–505.

Roger Mitton. 1996. *English spelling and the computer*. Longman Group.

Michael Mohler and Rada Mihalcea. 2009. Text-to-text semantic similarity for automatic short answer grading. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 567–575. Association for Computational Linguistics.

Jany Rademakers, Th J Ten Cate, and PR Bär. 2005. Progress testing with short answer questions. *Medical teacher*, 27(7):578–582.

Brian Riordan, Andrea Horbach, Aoife Cahill, Torsten Zesch, and Chungmin Lee. 2017. Investigating neural architectures for short answer scoring. In *Proceedings of the 12th Workshop on Innovative Use of NLP for Building Educational Applications*, pages 159–168.

Keisuke Sakaguchi, Michael Heilman, and Nitin Madnani. 2015. Effective feature integration for automated short answer scoring. In *Proceedings of the 2015 conference of the North American Chapter of the association for computational linguistics: Human language technologies*, pages 1049–1054.

Hanqing Tao, Shiwei Tong, Tong Xu, Qi Liu, and Enhong Chen. 2019. Chinese embedding via stroke and glyph information: A dual-channel view. *arXiv preprint arXiv:1906.04287*.

Hao-Chuan Wang, Chun-Yen Chang, and Tsai-Yen Li. 2008. Assessing creative problem-solving with automated text grading. *Computers & Education*, 51(4):1450–1466.

Ian H Witten, Eibe Frank, Leonard E Trigg, Mark A Hall, Geoffrey Holmes, and Sally Jo Cunningham. 1999. Weka: Practical machine learning tools and techniques with java implementations.

Shih-Hung Wu and Wen-Feng Shih. 2018. A short answer grading system in chinese by support vector approach. In *Proceedings of the 5th Workshop on Natural Language Processing Techniques for Educational Applications*, pages 125–129.

Shih-Hung Wu and Chun-Yu Yeh. 2019. A short answer grading system in chinese by cnn. In *2019 IEEE 10th International Conference on Awareness Science and Technology (iCAST)*, pages 1–5. IEEE.

Shaohua Yang, Hai Zhao, Xiaolin Wang, and Baoliang Lu. 2012. Spell checking for chinese. In *LREC*, pages 730–736.

Xi Yang, Yuwei Huang, Fuzhen Zhuang, Lishan Zhang, and Shengquan Yu. 2018. Automatic Chinese Short Answer Grading with Deep Autoencoder. In *Artificial Intelligence in Education*, pages 399–404, Cham. Springer International Publishing.

Xi Yang, Lishan Zhang, and Shengquan Yu. 2017. Can short answers to open response questions be auto-graded without a grading rubric? In *International Conference on Artificial Intelligence in Education*, pages 594–597. Springer.

Rongchao Yin, Quan Wang, Peng Li, Rui Li, and Bin Wang. 2016. Multi-granularity chinese word embedding. In *Proceedings of the 2016 conference on empirical methods in natural language processing*, pages 981–986.

Hongying Zan and Xuefeng Zhu. 2009. Nlp oriented studies on chinese functional words and the construction of their generalized knowledge base. *Contemporary Linguistics*, 2:124–135.

Torsten Zesch, Michael Heilman, and Aoife Cahill. 2015. Reducing annotation efforts in supervised short answer scoring. In *Proceedings of the Tenth Workshop on Innovative Use of NLP for Building Educational Applications*, pages 124–132.

Torsten Zesch and Andrea Horbach. 2018. Escrito-an nlp-enhanced educational scoring toolkit. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Hai Zhao, Deng Cai, Changning Huang, and Chunyu Kit. 2019. Chinese word segmentation: Another decade review (2007-2017). *arXiv preprint arXiv:1901.06079*.

Ramon Ziai, Niels Ott, and Detmar Meurers. 2012. Short answer assessment: Establishing links between research strands. In *Proceedings of the Seventh Workshop on Building Educational Applications Using NLP*, pages 190–200. Association for Computational Linguistics.