Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3013–3025 August 1–6, 2021. ©2021 ACL

Chinese L1 Linguistic Features

Category	Sub-category	ID	Linguistic Features
Shallow Features	Character complexity	1	percentage of most-common characters
		2	percentage of second-most-common characters
		3	percentage of all common-characters
		4	percentage of low-stroke-number characters
		5	percentage of medium-stroke-number characters
		6	percentage of high-stroke-number characters
		7	average number of strokes per word
	Word complexity	8	average number of characters per word
		9	average number of characters per unique word
		10	number of two-character words
		11	percentage of two-character words
		12	number of three-character words
		13	percentage of three-character words
		14	number of four-character words
		15	percentage of four-character words
		16	number of five-up-character words
		17	percentage of five-up-character word
	Sentence complexity	18	average number of multi-character words per sentence
		19	average number of words per sentence
		20	average number of characters per sentence
		21	average number of characters (including punctuations, numerical, and symbols) per sentence
	Document length	22	number of characters
		23	number of characters (including punctuations, numerical, and symbols)
Lexical Features	Adjectives	24	percentage of adjectives
		25	percentage of unique adjectives
		26	number of unique adjectives
		27	average number of adjectives per sentence
		28	average number of unique adjectives per sentence
	Functional words	29	percentage of functional words
		30	percentage of unique functional words as document vocabulary

Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3013–3025 August 1–6, 2021. ©2021 ACL

		31	number of unique functional words
		32	average number of functional words per sentence
		33	average number of unique functional words per sentence
	Verbs	34	percentage of verbs
		35	number of unique verbs
		36	percentage of unique verbs
		37	average number of verbs per sentence
		38	average number of unique verbs per sentence
	Nouns	39	percentage of general nouns
		40	number of unique general nouns
		41	percentage of unique general nouns
		42	average number of general nouns per sentence
		43	average number of unique general nouns per sentence
		44	percentage of All-Nouns
		45	number of unique All-Nouns
		46	percentage of unique All-Nouns
		47	average number of All-Nouns per sentence
		48	average number of unique All-Nouns per sentence
	Content words	49	percentage of content words
		50	number of unique content words
		51	percentage of unique content words
		52	average number of content words per sentence
		53	average number of unique content words per sentence
	Idioms	54	percentage of idioms
		55	number of unique idioms
		56	percentage of unique idioms
		57	average number of idioms per sentence
		58	average number of unique idioms per sentence
	Adverbs	59	percentage of adverbs
		60	percentage of unique adverbs
		61	number of unique adverbs
		62	average number of adverbs per sentence
		63	average number of unique adverbs per sentence
Syntactic Features	Phrases	64	average number of noun phrases per sentence
		65	average number of verbal phrases per sentence
		66	total number of noun phrases
		67	total number of verbal phrases

Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3013–3025 August 1–6, 2021. ©2021 ACL

		68	total number of prepositional phrases
		69	average length of noun phrases
		70	average length of verbal phrases
		71	average length of prepositional phrases
	Clauses	72	average number of sentences with clauses
		73	average number of sentences without clauses
		74	average number of clauses per sentence
		75	number of punctuation-clauses per sentence
		76	average number of words per punctuation-clause
	Sentences	77	average number of sentences
		78	average height of parse tree
		79	total number of dependency distances per sentence
		80	average number of dependency distances per sentence
Discourse Features	Entity Density	81	total number of entities
		82	total number of unique entities
		83	percentage of Entities
		84	percentage of unique entities
		85	average number of entities per sentence
		86	average number of unique entities per sentence
		87	percentage of named entities
		88	average number of named entities per sentence
		89	percentage of named entities against total number of entities
		90	percentage of Not-NE nouns
		91	average number of Not-NE nouns per sentence
		92	average number of Not-Entity nouns per sentence
	Cohesion	93	percentage of conjunctions against total number of words
		94	number of unique conjunctions
		95	percentage of unique conjunctions as document vocabulary
		96	average number of conjunctions per sentence
		97	average number of unique conjunctions per sentence
		98	percentage of pronouns
		99	number of unique pronouns
		100	percentage of unique pronouns
		101	average number of pronouns per sentence
		102	average number of unique pronouns per sentence

Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3013–3025 August 1–6, 2021. ©2021 ACL

Chinese L2 Linguistic Features

Category	Sub-category	ID	Linguistic Features
Shallow Features	Character Complexity	1	percentage of most-common characters
		2	percentage of second-most-common characters
		3	percentage of all common-characters
		4	percentage of low-stroke-number characters
		5	percentage of medium-stroke-number characters
		6	percentage of high-stroke-number characters
		7	average number of strokes per word
		8	percentage of HSK1 to HSK3-characters per sentence
		9	percentage of HSK4 to HSK5-characters per sentence
		10	percentage of HSK6-characters per sentence
		11	percentage of Not-HSK-characters per sentence
	Word Complexity	12	average number of characters per word
		13	average number of characters per unique word
		14	number of two-character words
		15	percentage of two-character words
		16	number of three-character words
		17	percentage of three-character words
		18	number of four-character words
		19	percentage of four-character words
		20	number of five-up-character words
		21	percentage of five-up-character word
		22	percentage of HSK1 to HSK3-words per sentence
		23	percentage of HSK4 to HSK5-words per sentence
		24	percentage of HSK6-words per sentence
		25	percentage of Not-HSK-words per sentence
	Sentence Complexity	26	average number of multi-character words per sentence
		27	average number of words per sentence
		28	average number of characters per sentence
		29	average number of characters (including punctuations, numerical, and symbols) per sentence
	Document Length	30	number of characters

Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3013–3025 August 1–6, 2021. ©2021 ACL

		31	number of characters (including punctuations, numerical, and symbols)
Lexical Features	Adjectives	32	percentage of adjectives
		33	percentage of unique adjectives
		34	number of unique adjectives
		35	average number of adjectives per sentence
		36	average number of unique adjectives per sentence
	Functional Words	37	percentage of functional words
		38	percentage of unique functional words as document vocabulary
		39	number of unique functional words
		40	average number of functional words per sentence
		41	average number of unique functional words per sentence
	Verbs	42	percentage of verbs
		43	number of unique verbs
		44	percentage of unique verbs
		45	average number of verbs per sentence
		46	average number of unique verbs per sentence
	Nouns	47	percentage of general nouns
		48	number of unique general nouns
		49	percentage of unique general nouns
		50	average number of general nouns per sentence
		51	average number of unique general nouns per sentence
		52	percentage of All-Nouns
		53	number of unique All-Nouns
		54	percentage of unique All-Nouns
		55	average number of All-Nouns per sentence
		56	average number of unique All-Nouns per sentence
	Content Words	57	percentage of content words
		58	number of unique content words
		59	percentage of unique content words
		60	average number of content words per sentence
		61	average number of unique content words per sentence
	Idioms	62	percentage of idioms
		63	number of unique idioms
		64	percentage of unique idioms
		65	average number of idioms per sentence
		66	average number of unique idioms per sentence

Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3013–3025 August 1–6, 2021. ©2021 ACL

	Adverbs	67	percentage of adverbs
		68	percentage of unique adverbs
		69	number of unique adverbs
		70	average number of adverbs per sentence
		71	average number of unique adverbs per sentence
Syntactic Features	Phrases	72	average number of noun phrases per sentence
		73	average number of verbal phrases per sentence
		74	total number of noun phrases
		75	total number of verbal phrases
		76	total number of prepositional phrases
		77	average length of noun phrases
		78	average length of verbal phrases
		79	average length of prepositional phrases
	Clauses	80	average number of sentences with clauses
		81	average number of sentences without clauses
		82	average number of clauses per sentence
		83	number of punctuation-clauses per sentence
		84	average number of words per punctuation-clause
	Sentences	85	average number of sentences
		86	average height of parse tree
		87	total number of dependency distances per sentence
		88	average number of dependency distances per sentence
		89	sentence complexity level
Discourse Features	Entity Density	90	total number of entities
		91	total number of unique entities
		92	percentage of Entities
		93	percentage of unique entities
		94	average number of entities per sentence
		95	average number of unique entities per sentence
		96	percentage of named entities
		97	average number of named entities per sentence
		98	percentage of named entities against total number of entities
		99	percentage of Not-NE nouns
		100	average number of Not-NE nouns per sentence
		101	average number of Not-Entity nouns per sentence
	Cohesion	102	percentage of conjunctions against total number of words
		103	number of unique conjunctions

		104	percentage of unique conjunctions as document vocabulary
		105	average number of conjunctions per sentence
		106	average number of unique conjunctions per sentence
		107	percentage of pronouns
		108	number of unique pronouns
		109	percentage of unique pronouns
		110	average number of pronouns per sentence
		111	average number of unique pronouns per sentence

English 33 Linguistic Features

Category	ID	Linguistic Features
Lexical Features	1	Lexical Density (LD)
	2	Type-Token Ratio (TTR)
	3	Corrected TTR
	4	Root TTR (RTTR)
	5	Bilogarithmic TTR (LogTTR)
	6	Uber Index (Uber)
	7	Lexical Word Variation (LV)
	8	Verb Variation-1 (VV1)
	9	Squared VV1 (SVV1)
	10	Corrected VV1 (CVV1)
	11	Verb Variation 2 (VV2)
	12	Noun Variation (NV)
	13	Adjective Variation (AdjV)
	14	Adverb Variation (AdvV)
	15	Modifier Variation (ModV)
	16	Proportion of words in AWL (AWL)
	17	Avg. Num. Characters per word (NumChar)
	18	Avg. Num. Syllables per word (NumSyll)
Syntactic Features	19	Mean length of a sentence
	20	average number of words per punctuation-clause
	21	number of punctuation-clauses per sentence
	22	average number of subordinate clauses per punctuation clause
	23	average number of subordinate clauses per sentence
	24	average number of co-ordinate phrases per punctuation clause

Learning Syntactic Dense Embedding with Correlation Graph for Automatic Readability Assessment

Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing, pages 3013–3025 August 1–6, 2021. ©2021 ACL

	25	average number of co-ordinate phrases per sentence
	26	average number of verb phrases per punctuation clause
	27	average number of noun phrases per sentence
	28	average number of verbal phrases per sentence
	29	average number of prepositional phrases per sentence
	30	average length of noun phrases
	31	average length of verbal phrases
	32	average length of prepositional phrases
	33	average height of parse tree