ООО «Когнитивные системы»
ОГРН: 1165029057490; ИНН: 5029214276
Адрес: 141014, Московская обл, г.Мытищи,
ул. Веры Волошиной дом 12, офис 714;
office@cogsys.company

Brain2Spell

Brain2Spell – это сервис, позволяющий распознавать слова и исправлять в них орфографические ошибки и опечатки. Являясь одним из модулей Большой Лингвистической Модели, алгоритм Brain2Spell нацелен на распознавание базовой составляющей языка – слова.

Сервис Brain2Spell функционирует на основе нейронной сети, обученной на корпусе в 120 тысяч слов. При этом проведенные тесты показали высокие результаты точности – 90% на тестовом наборе 25 тысяч слов с ошибками. Мы провели сравнение качества работы алгоритма с результатами, полученными на аналогичной задаче командами в ходе конкурса SpellRuEval-2016. Алгоритм Brain2Spell продемонстрировал лучшие показатели качества на словах с ошибками. Стоит отметить, что на текущий момент сервис распознает такие типы ошибок, как орфографические (кроме слитного написания слов и частиц), пропущенные буквы, лишние буквы и написанные в неправильном порядке буквы в словах.

ООО «Когнитивные системы»
ОГРН: 1165029057490; ИНН: 5029214276
Адрес: 141014, Московская обл, г.Мытищи,
ул. Веры Волошиной дом 12, офис 714;
office@cogsys.company

Group	Algorithm name or author	Words in corpus	Incorrect Words in sample	Accuracy/Recall		Year	Source
				Words with errors	All types		
English words	Dronen	120 000	4 349		0,92	2016	1
	Hodge, Austin	81 717	600		0,93	2003	3
	Wikipedia ≥ 10	42 622	1100	0,78		2013	9
	Wikipedia ≥ 8			0,75		2013	9
	Wikipedia ≥ 6			0,71		2013	9
	Wikipedia ≥ 4			0,68		2013	9
	Ahmad, Kondrak	580 000	508		0,79	2005	6
Russian words	**Brain2Word**	**118 637**	**1) 25 000 2) 660**	0,90	0,71*	**2017**	-
	MS Word 2007	-	56 000 Not publicly available	0,88		2015	8
	Yspell			0,88		2015	8
	Yandex			0,84		2015	8
	Google			0,80		2015	8
	CIGR Lemming			0,80		2015	8
	Aspell			0,70		2015	8
	MSU*	10 000 for tune up + dictionaries	1100 publicly available		0,70	2016	4
	Orfogrammatika*				0,61	2016	4
	MIPT*				0,58	2016	4
	ISP RAS*				0,56	2016	4
	HSE*				0,50	2016	4
	NLP@Cloud*				0,34	2016	5
	InfoQubes*				0,25	2016	4
	Dereza et al.	900 000	-		0,40	2016	8

Алгоритм был протестирован на двух наборах данных.

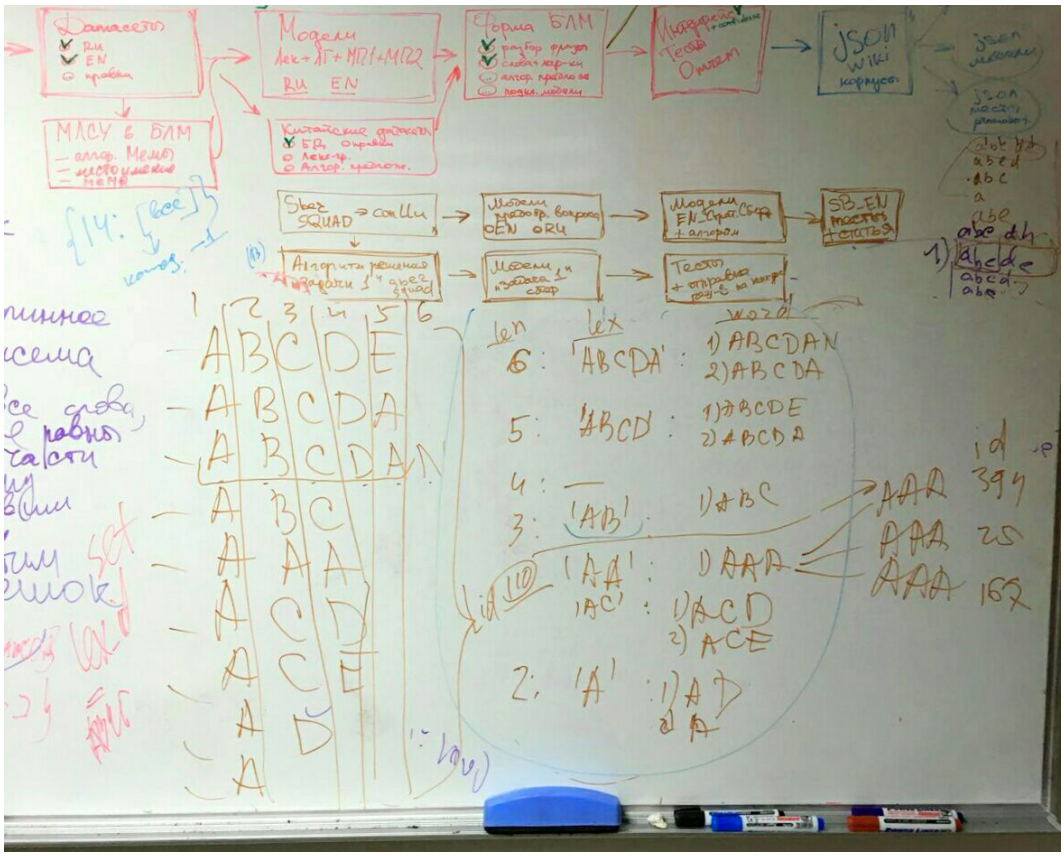
1) 25 тысяч слов были выбраны из набора из 60 тысяч слов, основанных на наиболее распространенных и актуальных ошибках написания слов - из исследования Карты Слова Д. Кулагина: Пересмотр подхода к составлению онлайновых словарей в постмобильную эпоху. Компьютерная лингвистика и интеллектуальные технологии: материалы международной конференции «Диалог 2017», 31 мая - 3 июня 2017 года. http://cogsys.company/data/test_25k_errs.csv

ООО «Когнитивные системы»
ОГРН: 1165029057490; ИНН: 5029214276
Адрес: 141014, Московская обл, г.Мытищи,
ул. Веры Волошиной дом 12, офис 714;
office@cogsys.company

2) Исходный набор из 1100 слов с ошибками содержал различные виды сгенерированных ошибок - два слова без пробелов, перемешанные буквы и т. д. Мы использовали набор данных из 1100 слов с ошибками - только орфографические ошибки и опечатки - из конкурса Spellrueval: Конкурс на автоматическую коррекцию правописания для русского языка. Сорокин А.А., Байтин А.В., Галинская И.Е., Шаврина Т.О. "Диалог-2016". 1-4 июня 2016 г.
http://cogsys.company/data/test_1200w_60-40.csv

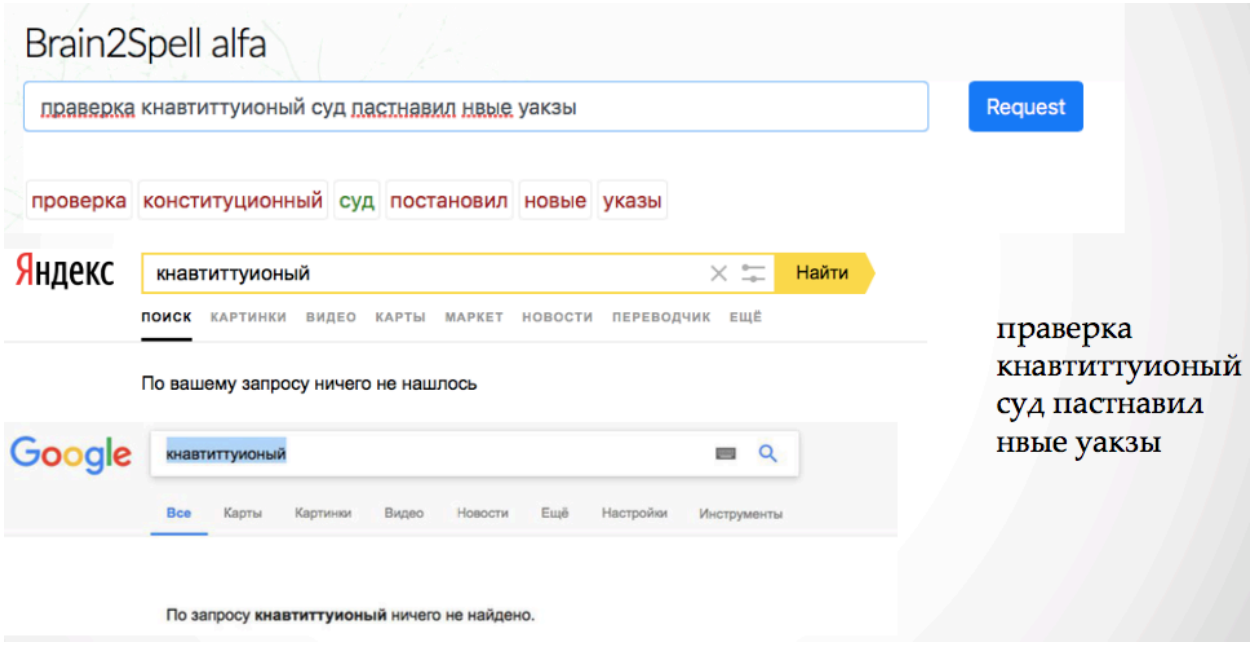
В ходе работы над алгоритмом были испробованы различные варианты сочетания признаков и классов модели для достижения лучшего результата.

ООО «Когнитивные системы»
ОГРН: 1165029057490; ИНН: 5029214276
Адрес: 141014, Московская обл, г.Мытищи,
ул. Веры Волошиной дом 12, офис 714;
office@cogsys.company



При этом сервис Brain2Spell лучше распознает неправильно написанные слова с большим числом ошибок, чем поисковые алгоритмы Яндекс и Google, которые не всегда могут распознать искаженные поисковые запросы.

ООО «Когнитивные системы»
ОГРН: 1165029057490; ИНН: 5029214276
Адрес: 141014, Московская обл, г.Мытищи,
ул. Веры Волошиной дом 12, офис 714;
office@cogsys.company



То же можно сказать и про систему исправления ошибок Microsoft Word, примеры работы которой в сравнении с Brain2Spell представлены в таблице ниже.

CORRECT	MISTAKE	MS_WORD	Brain2Spell
мужественный	мужиствиний	мужественней	мужественный
аплодисменты	обладесменты	обладесменты	аплодисменты
экскурсии	екскурся	экскурса	экскурсии
пятьдесят	педясят	подаст	пятьдесят
путешественник	путишествинек	путешественник	путешественник
препинания	припеннание	препинание	препинания
пассажирский	посожирсский	пассажирский	пассажирский
человек	ччеловввек	человеке	человек

Протестировать альфа-версию сервиса Brain2Spell можно по следующей ссылке - http://cogsys.company/ru/brain2spell.

Следующими ступенями создания Большой Лингвистической Модели будут создание алгоритма определения морфологических свойств слова, а также алгоритма синтаксического анализа предложений.