ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ПРОФЕССИОНАЛЬНОГО ОБРАЗОВАНИЯ

САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ

Факультет *Информационных Технологий и Программирования*
Направление (специальность) *Прикладная математика и информатика*
Квалификация (степень) *Бакалавр прикладной математики и информатики*
Кафедра *Компьютерных технологий* Группа *4538*

ПОЯСНИТЕЛЬНАЯ ЗАПИСКА

К ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЕ

Разработка классификатора для оценки качества перевода без использования эталонов перевода

Автор квалификационной работы *Попов А. Л.* (подпись)

Руководитель *Байтин А. В.* (подпись)

Консультанты:

а) По экономике и организации производства _____ (подпись)

б) По безопасности жизнедеятельности и экологии _____ (подпись)

в) _____ (подпись)

К защите допустить

Зав. кафедрой *Васильев В.Н.* (подпись)

«___» _____ 2014 г.

Санкт-Петербург, 2014 г.

Оглавление

Введение

Как исследователи, так и разработчики систем машинного перевода нуждаются в объективной, недорогой и качественной метрике качества перевода. Без оценки качества перевода невозможно судить как о абсолютном качестве перевода, так и разрабатывать и изменять систему машинного перевода.

Существует несколько решений, каждое из которых является в той или иной степени компромиссным. Во-первых, ручная оценка: эксперты (носители как языка исходного текста, так и языка перевода) оценивают переводы по определённым параметрам, например, грамотность и семантическая точность (адекватность) перевода. Этот способ считается эталонным, но при этом является самым дорогим и трудным (зачастую, найти носителей двух не самых распространненых языков крайне непросто).

Второй способ – автоматическая оценка качества перевода с использованием эталонов. В данном случае эксперты необходимы только на стадии подговки эталонов для тестового корпуса, а для сравнения перевода и эталона используют автоматические метрики. Разработка подобной метрики – непростая задача, именно поэтому этот способ также является компромисным – стоимость относительно ручной оценки меньше, но качество хуже. Но, зачастую, даже такой способ оказывается слишком дорогим для некоторого класса задач.

В рамках данной работы было необходимо внести вклад в разработку третьего, полностью автоматического способа оценки качества перевода – без использования каких-либо эталонов. Этот метод может быть достаточно груб и от него не требуется идеальной точности для успешного использования – он необходим там, где объемы тестовых данных и требования к скорости получения результатов не позволяют использовать другие методы.

В первой главе представлен краткий обзор предметной области. Во второй главе описаны предложеные изменения и описаны компоненты, которые в итоге были использованы в конечном решении, результаты работы которого описаны в последней главе.

Глава 1. Обзор предметной области

1.1. Перевод

В первую очередь, определим базовые понятия предметной области:

Определение 1.1. Перевод – процесс передачи смысла текста на исходном языке в виде эквивалентного по смыслу текста на другом языке (языке перевода).

В рамках данной работы нам будут интересны следующие виды перевода:

1. Ручной перевод
2. Статистический машинный перевод
3. Машинный перевод на основе правил
4. Гибридный машинный перевод

Определение 1.2. Ручной перевод – перевод, полученный от эксперта-переводчика. Соответственно, машинный перевод – перевод, выполненный с помощью програмного обеспечения.

Определение 1.3. Статистический машинный перевод – парадигма машинного перевода, где перевод порождается с помощью статистических моделей, которые извлекаются с помощью анализа паралельных двуязычных корпусов текста.

Статистический перевод в данной работе представляет наибольшую ценность, поэтому он будет описан более подробней.

Определение 1.4. Машинный перевод на основе правил – парадигма машинного перевода, где перевод порождается с помощью лингвистической информации о исходном языке и языке перевода, получаемой, в основном

из словарей и грамматик, покрывающих основные семантические, морфологические и синтаксические особенности языков. Грамматики в данном случае составляются людьми вручную.

Определение 1.5. Гибридный машинный перевод – парадигма машинного перевода, где перевод порождается с помощью совокупности двух предыдущих парадигм.

1.1.1. Статистический машинный перевод

Статистический машинный перевод может быть создан на основе различных моделей языка и перевода. В рамках данной работы нас интересуют фразовые модели перевода. Для полноты повествования, основные идеи фразовых моделей перевода описаны в данном параграфе.[1]

Опишем идею стандартной фразовой модели перевода. Исходный текст разбивается на так называемые фразы (любая последовательность из нескольких слов). Затем, каждая фраза независимо переводится на язык перевода. Наконец, полученые фразы необходимо упорядочить (порядок фраз исходного и конечного текста может не совпадать). Для того, чтобы выбрать подходящий перевод, необходимо фразовая таблица переводов, где каждой фразе будут сопоставлены возможные переводы и их вероятности. Модель, основанная на фразах, а не словах, обладает несколькими преимуществами. Во-первых, слова вряд ли можно считать лучшей атомарной единицей перевода (как минимум, из-за частых отображений один-ко многим). Во-вторых, перевод группы слов вместо конкретного слова помогает разрешать неоднозначности благодаря наличию контекста.

За порядок фраз отвечает специальная модель упорядочивания, основанная на расстоянии. Достаточно знать, что идея стоящая за подобной моделью проста – перемещение фраз на большие дистанции менее вероятно, нежели перемещение на небольшую дистанцию.

Наконец, после получения некоторого предложения из перестановки переведённых фраз, необходимо умение определить, насколько вероятно

присутствие такого предложения в языке перевода. Для этого для гипотезы перевода считается оценка по языковой модели. Наиболее общепринятый подход для моделирования языка – использование модели языка построеной по N-граммам.

Итого, в простой фразовой модели перевода, перевод будет выбираться с помощью следующего выражения:

$$e_{best} = argmax_e \prod_{i=1}^{I} \phi(\bar{f}_i|\bar{e_i}) d(start_i - end_{i-1} - 1) \prod_{i=1}^{|e|} p_{LM}(e_i|e_1 \dots e_{i-1}) \quad (1.1)$$

Где, e_{best} – лучший перевод, ϕ – вероятность перевода фразы f_i в e_i, полученная из фразовой таблице переводов, d – вероятность перемещения фразы на определённое расстояние, полученная из модели упорядочивания, а p_LM – вероятность принадлежности предложения языку перевода. Исходя из этого, определим необходимые для нас в дальнейшем две величины:

Определение 1.6. LM-score – вероятность принадлежности некоторого предложения e некоторому языку.

$$\text{LM} - \text{score} = p_{LM}(e_i|e_1 \dots ei - 1) \quad (1.2)$$

Определение 1.7. PT-score – вероятность того, что некоторое предложение e является переводом f, полученная из фразовой модели перевода.

$$\text{PT} - \text{score} = max \prod_{i=1}^{Ie|} \phi(\bar{f}_i|\bar{e_i}) d(start_i - end_{i-1} - 1) \quad (1.3)$$

1.2. Оценка качества перевода

Вопрос «Насколько хорош машинный перевод в наше время?» довольно непрост, так как для любого текста существует множество различных, в равной степени корректных переводов, а не только один, единственно верный ответ. Также перевод может быть корректен граматически, но абсолютно бессмысленнен, либо, наоборот, общий смысл понятен, но, при этом, согласование и грамматическая структура сильно нарушена. Также, например, на практике крайне сложно установить, было ли пропущено

важное отрицание в предложении, что, естественно, сильно меняет семантику предложения. Существует несколько подходов для оценки качества перевода.

1.2.1. Ручная оценка качества перевода

Определение 1.8. Ручная оценка — способ оценки машинного перевода с помощью человеческих сил, когда асессоров просят оценить качество конкретного перевода по определённым критериям и определённой шкале, например, грамотность и адекватность перевода.

Для ручной оценки требуется асессоры, владеющие двумя языками сразу. Таких людей трудно найти и нанять, особенно для оценки непопулярных направлений перевода. Кроме того, таким образом можно разметить лишь относительно небольшие наборы. Есть более распространенный и более дешевый вариант ручной оценки, когда переводчик переводит предложения из тестового корпуса (подготовка эталонов перевода) и, затем, асессор владеющий языком конечного перевода сравнивает машинный перевод с эталонным. Этот подход, хоть и проще, остается довольно дорогим для разработки. Итого, ручная оценка качества слишком дорога, требовательна к асессорам и, зачастую, довольно несогласованна (оценка больше зависит от мнения асессора, нежели объективного качества). С другой стороны, несогласованность обычно решается изменением метрики и все качество любых автоматических метрики проверяют корреляцией с ручной оценкой, которая считается эталонной.

1.2.2. Автоматическая оценка качества перевода с использованием эталонов

Для упрощения и удешевления оценки качества, используются различные автоматические метрики[1], которые работают при наличии одного или нескольких эталонов перевода — Точность/Полнота, WER, BLEU, NIST, METEOR. Автоматические метрики также требуют затрат на раз-

метку тестовых данных, более того, они работают лучше с увеличением количества эталонов (до определённого порога, дальше уже возможна вероятность случайного хорошего совпадения). У подобных метрик существуют свои недостатки — абсолютные показатели трудно интерпретировать, значения метрик не несут никакого интуитивно понятного смысла. Но главная концептуальная проблема заключается в том, что в основе подобных метрик стоит сравнение текста перевода и эталона на основе формальных факторов (например, количество совпавших n-грамм). Автоматические метрики с трудом учитывают, передал ли перевод смысловое содержание, не исказил ли перевод важные части перевода.

Рассмотрим одну из самых распространенных метрик – BLEU. Идея метрики состоит в подсчете точности n-грамм между эталоном и переводом некого текста. Например, вариант метрики с ограничением до 4-грамм выглядит следующим образом:

$$\text{Bleu} - 4 = min(1, \frac{\text{output} - \text{length}}{\text{reference} - \text{length}}) \sqrt[4]{\prod_{i=1}^{4} precision_i} \tag{1.4}$$

где $precision_i$ – отношение количества корректных i-грамм к общему количеству i-грамм в переводе.

Лучше всего такая метрика работает не на уровне предложений, а на уровне большого текста (проблема на маленьком объеме текста такова, что метрика зачастую обнуляется из-за отсутствия совпадающих 4-грамм), но есть вариант метрики, который подходит для сравнения на уровне предложения.

Определение 1.9. Smoothed Bleu или BleuS[2] — видоизмененая метрика Bleu для подсчета на уровне предложений. Для того, чтобы получить BleuS, нужно к количеству корректных и к общему количеству n-грамм добавить единицу для всех $n > 1$

Таким образом, мы оптимизируем метрику под уровень предложения и, при этом, значение метрики все-равно будет равняться нулю тогда

и только тогда, когда ни одно слово из перевода не совпадет ни с одним словом из эталона.

Еще один вариант улучшения данной метрики – использовать несколько эталонов перевода. Вообще, подобные метрики довольно неплохо корреллируют с ручными оценками в статистическом машинном перевода, но при этом имеют проблемы с другими видами перевода. [1]

1.2.3. Автоматическая оценка качества перевода без использования эталонов

Интересным направлением в области оценки качества перевода является попытка отказаться от использования эталонов. Это изменение делает оценку качества практически бесплатной. Получение подобного инстумента могло было бы открыть следующие возможности:[3]

1. Принятие решения, достаточно ли хорош перевод для публикации без постредактирования
2. Информирование читателя перевода о качестве перевода
3. Фильтрация предложений, которые не имеет смысла редактировать, а лучше перевести при помощи профессионального переводчика
4. Выбор лучшего перевода из нескольких вариантов

Актуальность данной задачи доказывает то, что созданием подобной метрики занимаются в рамках семинара по статистическому машинному переводу с 2012 года.[4][3] Каждый год формулировка конкретной задачи меняется, переводы пытаются оценивать как просто с помощью абстракной оценки (1 – перевод идеальный, 2 - одна-две ошибки, 3 – ужасный), так и с помощью прикладных метрик (HTER – сколько шагов потребуется сделать переводчику, чтобы обработать полученный перевод для публикации[3]).

Традиционным подходом к решению данной задачи является обучение классификатора или регрессии с помощью методов машинного обучения.[4]

1.3. Машинное обучение

Определение 1.10. Машинное обучение – подраздел искуственного интеллекта, изучающий модели, способные обучаться из данных и алгоритмы для их построения и обучения.

Определение 1.11. Классификация – задача, в которой существует множество объектов и классы, к которым объекты могут принадлежать. Необходимо по данному объекту определить его класс.

Определение 1.12. Регрессия – задача, в которой существует множество объектов и переменная, которая зависит от объекта (зависимость неизвестна). Необходимо по данному объекту вычислить зависимую переменную.

В данном разделе мы рассмотрим методы машинного обучения, которые будут использованы в данной работе.

1.3.1. SVM (Метод опорных векторов)

SVM (Support Vector Machines) – алгоритм машинного обучения, используемый для решения задач классификации. Опишем общую идею алгоритма:

Пусть у нас есть объекты, описанные как точки в некотором пространстве R^n. Для точек из обучающего множества нам известно, к какому классу они принадлежат. Допустим, выборка линейно-разделима (можно построить такую разделяющую гиперплоскость, что по одну сторону от неё будут объекты одного класса, а другого класса – по другую. Тогда, данный алгоритм будет стремится построить такую разделяющую гиперплоскость, что минимальный отступ (geometric margin) от неё до всех объектов был максимальным.

1.3.2. SVR (Support Vector Regression)

Определение 1.13. SVR (Support Vector Regression)[5] – разновидность метода Support Vector Machines, которая применима для задач регрессии.

1.3.3. Gradient Boosting Trees

Gradient Boosting[6] – специальная техника машинного обучения, которая строит предсказывающую модель в форме ансамбля слабых моделей. Обычно, Gradient Boosting основывается на деревьях решений фиксированного размера в качестве базовых моделей.

1.3.4. MatrixNet

MatrixNet[7] – это проприетарная реализация Gradient Boosting Trees, которая в некоторых моментах отличается от реализации автора метода, TreeNet[8].

MatrixNet используется в качестве стандартного алгоритма машинного обучения в Яндексе для широкого спектра задач. Этот алгоритм можно использовать для задач регрессии, классификации, мультиклассификации, а также, главным образом, для задач ранжирования. Ранжирование осуществляется с помощью оптимизации по метрике pFound[9].

$$\text{pFound} = \sum_{i=1}^{n} pLook(i) \cdot pRel(i) \tag{1.5}$$

$$\text{pLook}(i) = pLook(i-1)\dot{(}1 - pRel(i-1))\dot{(}1 - pBreak) \tag{1.6}$$

1.4. Лингвистика

Для извлечения некоторых факторов, напрямую относящимся к различным лингвистическим признакам и свойствам, необходимо использовать некоторое количество лингвистических инструментов:

1.4.1. Частеречная разметка

Определение 1.14. Частеречная разметка – процесс разметки слов в тексте частями речи, основанный как на самих словах, так и на их контексте в тексте.

Для разметки предложений будет использоваться TreeTagger[10] – специальный, языконезависимый инструмент, использующийся для частеречной разметки.

1.4.2. Синтаксический анализ

Определение 1.15. Синтаксический анализ – процесс анализа грамматической структуры текста на естественном языке.

По последовательности слов, программа-анализатор способна определить отношения в тексте, которые соответсвуют какому-либо грамматическому формализму (например, подлежащие, сказуемые, объекты и субъекты действий). В данной работе будут фигурировать Stanford Parser[11] и Berkeley Parser[12].

1.5. Постановка задачи

Необходимо разработать метод оценки качества перевода в реальном времени, без использования эталонов в следующем виде[13]: Дано: Набор исходных предложений и их переводов, полученных из разных источников. Необходимо для каждого перевода оценить его качество, присвоив ему численную метку от 1 до 3, где:

- 1 – идеальный перевод, не требуется пост-обработка
- 2 – перевод с 2-3 ошибками и, возможно, дополнительными ошибками, которые легко исправить (пунктуация, регистр букв)
- 3 – очень плохое качество перевода, нельзя легко исправить.

Также, все переводы необходимо отранжировать путем присваивания индекса от 1 до N (где 1 – перевод высшего качества, N – перевод худшего качества)

1.5.1. Корпус данных

Для обучения алгоритма представлен корпус данных, состоящий из 954 предложений на английском языке, каждое из которых имеет по 4 перевода на испанский язык из разных источников (ручной, статистический машинный, машинный на основе правил, гибридный). Итого, 3814 испанских переводов. Каждый из переводов размечен экспертами и содержит эталонную численную метку качества. Для тестирования предоставлен корпус данных, собраный таким же образом, только для тестирования используется 150 исходных предложений на английском (соответственно, 600 переводов на испанский).

1.5.2. QuEst

QuEst[14] – специальная утилита для оценки качества перевода без использования эталонов. Она позволяет собирать множество различных факторов из исходных текстов предложений и, также, с помощью этой утилиты можно построить классификатор или регрессию по собранным факторам. Решение, основанное на этой утилите считается baseline решением, и главная задача – улучшить это решение по различным ключевым метрикам. Baseline решение основано на SVR, использующее ядро Гаусса, построенное на 17 факторах:

- количество токенов в исходном предложении
- количество токенов в переводе
- средняя длина токена в исходном предложении
- Вероятность принадлежности исходного предложения модели языка
- Вероятность принадлежности перевода модели языка
- Среднее количество вхождений каждого слова из перевода в переводе.

- Среднее количество вариантов перевода по словам из исходного предложения (взятое из таблицы переводов по модели IBM 1, при этом вероятность перевода > 0.2)
- Среднее количество вариантов перевода по словам из исходного предложения (взятое из таблицы переводов по модели IBM 1, при этом вероятность перевода > 0.01) взвешенное по обратной частоте слова в корпусе изначального языка.
- Процент слов в первом квантиле частотности слов из исходного предложения в корпусе изначального языка.
- Процент слов в четвертом квантиле частотности слов из исходного предложения в корпусе изначального языка.
- Процент биграмм в первом квантиле частотности слов из исходного предложения в корпусе изначального языка.
- Процент биграмм в четвертом квантиле частотности слов из исходного предложения в корпусе изначального языка.
- Процент триграмм в первом квантиле частотности слов из исходного предложения в корпусе изначального языка.
- Процент триграмм в четвертом квантиле частотности слов из исходного предложения в корпусе изначального языка.
- Процент слов из исходного предложения, присутствующих в тренировочном корпусе статистического машинного перевода
- Количество знаков препинания в исходном предложении
- Количество знаков препинания в конечном предложении

Глава 2. Теоретическое исследование

В данной главе описываются главные компоненты решения задачи: факторы и методы машинного обучения, с помощью которых создавались варианты решения.

2.1. Факторы

Базисом набора факторов стал набор из 79 Black Box факторов, которые умеет извлекать QuEst[15]. Хотелось бы заметить, что в этом наборе множество действительно уникальных факторов меньше, чем множество всех факторов: многие факторы являются лишь вариантами одной и той же идеи с разными параметрами. Набор из 17 baseline факторов является подмножеством вышеупомянутого набора.

Проблема факторов, связанных с моделями языка, которые извлекаются с помощью QuEst состоит в том, что они построены на небольшой, ограниченной модели языка, построенной по небольшому корпусу текстов. Вычисление фактора вероятности существования предложения в языковой модели, построенного на более крупной и полной языковой модели должно давать лучшие результаты. Именно из этих суждений были добавлены факторы LM-score исходного предложения и перевода по продуктовой языковой модели, использующейся в переводном сервисе компании «Яндекс».

В рамках данной задачи при разработке нет доступа к эталонам перевода. Но возможно интерпретировать подобным образом ответ какого-либо независимого машинного перевода. Это будет не точный эталон, но при улучшении качества машинного перевода будет расти и надежность данной фичи. Назовём такой ответ псевдоэталоном. Тогда можно посчитать стандартную метрику качества перевода с использованием эталона, например, BleuS. Таким образом мы получили меру точности совпадения гипотезы перевода и псевдоэталона. [16]

2.1.1. Факторы сложности исходного предложения

Множество факторов замеряют характеристики исходного предложения. Можно заметить, что по отдельности, в отрыве от факторов, относящимся к переводу, эти факторы замеряют некоторую меру «сложности» исходного предложения. Интуиция подсказывает, что, чем сложнее предложение, тем больше вероятность ошибки перевода. В рамках эксперимента предлагается попробовать еще несколько факторов, осваивающих это направление:

- Количество and во входном предложении.
- Количество вопросительных слов на wh во входном предложении.
- Количество местоимений во входном предложении.
- Количество вспомогательных глаголов во входном предложении.
- Отношение количества существительных к количеству предлогов во входном предложении.
- Количество глаголов во входном предложении.
- Отношение количества глаголов к количеству союзов во входном предложении.
- Количество служебных слов во входном предложении.

Для извлечения этих факторов потребовалась частеречная разметка.

2.1.2. Факторы, использующие другой подход к оценке длин

Факторы, отвечающие за отношение длин исходных и конечных предложений входят в baseline набор и показывают неплохие результаты. Попробуем усовершенствовать эти факторы, немного изменив способ построение и сравнения длин (в исходных факторах в качестве длины предложения используют количество токенов):

- Длина входного предложения в словах.

- Длина перевода в словах.
- Количество знаков препинания во входном предложении.
- Количество знаков препинания в переводе.
- Отношение длин исходного предложения и перевода в словах.

2.1.3. Факторы, основанные на синтаксическом анализе

Развить направления сложности исходного предложения и оценить качество конечного предложения возможно могут помочь синтаксические факторы:

- Вероятность разбора входного предложения парсером Stanford.
- Вероятность разбора входного предложения парсером Stanford, деленная на длину предложения.
- Вероятность разбора перевода парсером Berkley.
- Вероятность разбора входного предложения парсером Berkley.
- Количество вариантов разбора для входного предложения (парсер Berkley).

2.1.4. Морфологическая и синтаксическая схожесть

Также для экспериментальной проверки были предложены факторы, отвечающие за морфологическую и синтаксическую схожесть исходного предложения и перевода.

- Отношение числа содержательных глаголов.
- Отношение числа вспомогательных глаголов.
- Отношение числа пассивных форм.
- Отношение числа модальных глаголов.
- Разность количества основ.
- Количество глаголов без субъекта во входном предложении.

2.1.5. Человечность перевода

Также из предложенных идей для факторов выделялась идея оценить «Человечность» перевода – насколько по грамматической структуре перевод похож на человеческий и не похож на машинный. Уже существовали попытки отличить машинный перевод низкого качества от человеческого перевода, возможно, применив использованные методы для обучения фактора можно будет получить неплохой результат[17].

2.2. Машинное обучение

Помимо факторов, необходимо обозначить алгоритмы машинного обучения, которые будут использованы для эксперимента и построения решения.

2.2.1. Линейная регрессия

Предположительно, линейная регрессия в данной задаче вряд ли даст какие-либо качественные результаты относительно SVR по многим причинам, например: множество фич сомнительного качества, вероятное отсутствие какой-либо линейной зависимости.

2.2.2. SVR

Support Vector Regression является классическим выбором[3] исследователей для решения задач в области оценки качества перевода без использования эталона. Недостатком данного алгоритма является то, что он крайне подвержен переобучению, если присутствуют плохие, шумные, нерабочие факторы. Также SVR не очень хорошо работает с большими наборами факторов и заметно улучшает результаты при грамотном отборе факторов. Все вышеперечисленные ньюансы сильно касаются нашего решения (в силе множества факторов нет уверенности, самих факторов большое количество), поэтому необходимо проводить отбор. В качестве реали-

зации в рамках данной работы была взята реализация SVR из scikit-learn, пакета для языка Python, предназначенного для машинного обучения. В нём, в свою очередь, используется svm-light: реализация алгоритма на C.

2.2.3. Random Forest

Данный алгоритм, основанный на ансамблях деревьях решений, использовался в качестве алгоритма машинного обучения на данной задаче и показал удовлетворительные результаты. У нас есть возможность сравнить его работу с другими алгоритмами, и, в частности, с MatrixNet, который тоже, в своей основе, является ансамблем деревьев решений.

2.2.4. MatrixNet

Если судить по отчетам с семинаров прошлых лет, до этого в работах исследователи не использовали алгоритм Gradient Boosting Decision Trees и, тем более, не имели доступа к его проприетарной реализации компании «Яндекс» MatrixNet. Именно поэтому этот алгоритм представляет большой интерес и может показать какие-то до этого неизвестные результаты. Данный алгоритм устойчив к переобучению, поэтому он не так остро нуждается в отборе факторов (в отличии от SVR). Также MatrixNet способен оценивать факторы по «эффективности», что само по себе является интересным способом проводить отбор факторов и их анализ. MatrixNet способен работать в разных режимах, например, он способен решать задачу регрессии или задачу ранжирования по разработанной в «Яндексе» метрике PFound. Если с режимом регрессии всё ясно – можно просто воспользоваться им «из коробки», то режим ранжирования (а это естественная специализация этого алгоритма) можно будет использовать только вместе с определёнными модификациями.

2.2.5. MatrixNet: ранжирование по pFound

Как можно воспользоваться инструментом для ранжирования поисковой выдачи в рамках нашей задачи? Допустим, что исходное предложение – аналог поискового запроса в мире оценки качества перевода. Соответственно, документами будут возможные переводы для данного предложения, и MatrixNet будет заниматься ранжированием этих переводов по качеству. В качестве входных данных, необходимо разметить переводы вместо меток качества значениями релевантности для метрики pFound – вероятностью того, что перевод устроит получателя. Допустим, что метку качества «1» можно считать удовлетворительным переводом с вероятностью 0.9, метку качества «2» с вероятностью 0.5, а перевод с оценкой «3» никого не устроит, поэтому вероятность удовлетворительности перевода будет считаться за нулевую.

Вопрос с ранжированием более или менее ясен – как же с помощью этого метода решить задачу классификации или регрессии? Значения релевантности документов в MatrixNet не являются надежной абсолютной оценкой сами по себе, они показыют лишь значение, необходимое для сравнения между документами для одного запроса. Для решения этой проблемы можно воспользоваться SoftMax преобразованием – обобщением логистической функции, которое преобразовывает набор вещественных величин в распределенние вероятностей, переводя каждое число из набора в величину принадлежащую отрезку $[0, 1]$.

$$\sigma(\mathbf{z})_j = \frac{e^{\mathbf{z}_j}}{\sum_{k=1}^{K} e^{\mathbf{z}_k}} \tag{2.1}$$

Далее, полученные значения можно будет переводить в необходимые нам классы с помощью каких-либо порогов (допустим, будем считать, что если величина преобразованного элемента < 0.2, то соответствующему переводу нужно поставить метку качества «1», и. т. д.). Обучения конкретным границам выполним с помощью обучающего множества.

Подобный подход позволил коллегам улучшить качество классифи-

катора, определяющего, является ли данный опечаточный запрос автозаменой или нет (необходимо ли сразу менять выдачу по опечаточному запросу на исправленный, или же нужно спросить пользователя об опечатке).

2.2.6. Отбор факторов

В данной работе есть две причины, которые мотивируют проводить отбор факторов – факторов огромное количество и их работоспособность в большинстве своём непроверена и стоит под большим вопросом. С одной стороны у нас есть возможность отбирать факторы по «Эффективности» из прогонов MatrixNet, с другой можно воспользоваться простым методом жадного отбора факторов – на каждом шаге отбора будем оставлять фактор, дающий наибольший рост ключевых метрик на валидационном множестве.

Глава 3. Практические исследования

В данной главе описываются проведенные эксперименты и ключевые результаты. Полученные решения сравниваются с Baseline решением по ключевым метрикам.

3.1. Метрики оценки качества классификатора

Описанные ниже метрики используются для сравнения решений данной задачи. Именно эти метрики используются организаторами WMT2014[13] для оценки отправленных решений. Метрики делятся на два типа: те, что оценивают задачу численной оценки переводов и те, что оценивают ранжирование всех переводов по качеству.

3.1.1. MAE

Метрика, оценивающая абсолютную среднюю ошибку между предсказанной и реальной оценкой качества переводов на тестовом множестве. Чем меньше ошибка, тем лучше классификатор.

$$\text{MAE} = \frac{1}{n}\sum_{i=1}^{n}|\hat{y}_t - y_i| \tag{3.1}$$

3.1.2. RMSE

Метрика, оценивающая корень среднеквадратичной ошибки между предсказанной и реальной оценкой качества переводов на тестовом множестве. От MAE отличается тем, что штрафует большие ошибки сильнее.

$$\text{RMSE} = \sqrt{\frac{\sum_{t=1}^{n}(\hat{y}_t - y_t)^2}{n}} \tag{3.2}$$

•

3.1.3. DeltaAvg

Эта метрика была предложена в 2012 году на семинаре по машинному переводу WMT2012[4] для оценки задания из данной области с точки зрения ранжирования. Опишем способ подсчета метрики:

Для данного перевода s, $V(s)$ представляет из себя функцию, которая сопоставляет некую внешнее значение данному переводу. Тогда расширим определение на множество переводов S – $V(S) = average(V(s)), s \in S$. Тогда для ранжированного множества S и параметра n определим S_i как i-тый квантиль множества S. $S_{i,j} = \bigcup_{k=i}^{j} S_k$. Тогда:

$$\mathrm{DeltaAvg_V}[n] = \frac{\sum_{k=1}^{n-1} V(S_{1,k})}{n-1} - V(S) \tag{3.3}$$

• И, наконец, опишем определение самой метрики:

$$\mathrm{DeltaAvg_V} = \frac{\sum_{n=2}^{|S|/2} \mathrm{DeltaAvg_V}[n]}{|S|/2 - 1} \tag{3.4}$$

3.1.4. Коэффициент ранговой корреляции Спирмена

Коэффициэнт ранговой корреляции Спирмена задается как коэффициент корреляции Пирсона между ранговыми переменными. В данном случае, каждому значению качества перевода присваивается ранк относительно качества других переводов.

$$\rho = \frac{\sum_i (x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_i (x_i - \bar{x})^2 \sum_i (y_i - \bar{y})^2}} \tag{3.5}$$

3.2. Оценка качества предложенных факторов

Все предложенные группы факторов будем по очереди подмешивать к baseline факторам и замерять изменение качества на отдельном тестовом множестве. Заметим то, что эта цифра будет замеряться на SVR и будет справедлива толька для линейных моделей. В другом эксперименте, мы будем использовать нелинейную модель, основанную на деревьях решений

(MatrixNet) и качество данных факторов будет меряться по другой технике. Заметим также, что для финального решения нам всё-равно будет необходимо провести отбор факторов с целью построения максимально целостной модели с наименьшим количеством плохо работающих факторов.

Группа факторов	MAE	RMSE
BASELINE	0.497903	0.635140
Сложность	0.496720	0.636693
Другой подход к оценке длин	0.496649	0.634287
Модель языка Яндекса	0.487268	0.614696
Синтаксические факторы	0.494234	0.630582
Псевдоэталон	0.495454	0.633100

Таблица 3.1 Влияние подмешивания новых групп факторов к baseline факторам на результат

Все факторы в той или иной степени улучшают качество, что видно по значениям целевых метрик. Серым цветом выделены статистически значимые улучшения. Лучше всего проявил себя фактор вероятности принадлежности перевода в рамках модели языка Яндекса. Скорей всего, это просто следствие более большой и точной модели.

3.2.1. Человечность перевода

Если реализовать фактор человечности перевода в виде «нечестного» фактора, который умеет с идеальной точностью определять, перевод человеческий или нет, то RMSE улучшается на 10

К сожалению, идеи из статьи, посвященной классификации машинного перевода плохо себя показали в эксперименте. Факторы, которые использовались в статье (присутствие специфических синтаксических конструкций) слабо проявили себя в обучающем множестве. Скорей всего это связано сразу с несколькими моментами:

- Статья была написана в предположении, что мы отличаем статистический машинный перевод от человеческого, тогда как у нас в обучающем множестве помимо статистического перевода присутствуют

машинный перевод на основе правил и гибридный машинный перевод.

- Факторы, описанные в статье должны неплохо работать на уровне текста, но в масштабе нашего обучающего набора они себя не смогли проявить.

3.3. Отбор факторов

Для отбора факторов, напомним, использовались два способа:

3.3.1. Жадный отбор факторов

На каждом шаге в текущее множество факторов добавлялся фактор, дающий наибольший вклад в качество классификации. Качество замерялось с помощью 5-fold кросс-валидации на обучающем множестве.

Таким образом было отобрано 5 факторов (дальнейшие факторы либо делали слишком малый вклад в качество, либо неоднозначно влияли на качество обучающего относительно тестового множества.

Минимальное RMSE по обучающему множеству	Минимальное RMSE по тестовому множеству	Выбранная метрика
0.647	0.648	Yandex LM-score
0.613	0.610	Количество токенов в переводе
0.606	0.602	Псевдоэталон
0.604	0.593	QuEst LM-score
0.601	0.588	Количество существительных в исходном предложении

Таблица 3.2 Результаты отбора факторов

3.3.2. «Эффективность» факторов в MatrixNet

В данной таблице представлены уникальные по смыслу факторы (группы одинаковых по смыслу но разных по параметру факторов замене-

ны лучшим фактором из группы), отсортированные MatrixNet'ом по параметру «Эффективности».

Фактор	Эффективность
Yandex LM-score перевода	1
Абсолютная разница «;» между исходным предложением и переводом, нормализованная по длине перевода	0.8653049923
Абсолютная разница «,» между исходным предложением и переводом, нормализованная по длине перевода	0.4338843796
Perplexity перевода в языковой модели QuEst	0.2004979948
Псевдоэталон	0.1592566087
Количество токенов в переводе	0.09388195087
Вероятность разбора перевода Berkley парсером	0.08120818683
Отношение количества токенов в переводе и исходном предложении	0.07415779046
Вероятность разбора исходного предложения Berkley парсером	0.05957234285
Среднее количество переводов на слово в исходном предложении	0.04499589301

Таблица 3.3 Результаты отбора факторов по «эффективности»

3.4. Результаты

Итого, полученные значения ключевых метрик на разных решениях продемонстрированны в таблице 3.4:

Решение	**MAE ↓**	**RMSE ↓**	**DeltaAvg ↑**	**Spearman Corr ↑**
MATRIXNET (106 факторов)	0.50	**0.62**	0.22	0.35
MATRIXNET (10 факторов)	0.50	**0.62**	0.24	0.35
MATRIXNET PFOUND	**0.40**	0.64	**0.28**	**0.44**
SVR (Жадный отбор)	0.49	0.63	0.19	0.30
BASELINE (QUEST)	0.51	0.64	0.17	0.25
TOP 1 WMT2014 (RTM-PLS-TREE)	0.49	0.61	0.26	0.41

Таблица 3.4 Общее сравнение результатов

- MATRIXNET (106 факторов) – решение, использующее все факторы и MatrixNet в режиме регрессии.
- MATRIXNET (10 факторов) – решение, использующее лучшие по эффективности факторы и MatrixNet в режиме регрессии.
- MATRIXNET PFOUND – решение, использующее все факторы и MatrixNet в режиме ранжирования по PFound.
- SVR (Жадный отбор) – решение на SVR, использующее факторы, полученные в результате жадного отбора факторов.
- BASELINE (QUEST) – решение на SVR, использующее 17 baseline BlackBox факторов, которое необходимо было улучшить
- TOP 1 WMT2014 (RTM-PLS-TREE) – лучшее решение, полученное в ходе решения совместной задачи в рамках WMT2014 паралельно с этой работой.

3.5. Анализ полученного решения

Все полученые решения статистически значимо лучше BASELINE решения. Решение MATRIXNET PFOUND показало лучшие результаты по всем метрикам и, в целом, было показано хорошее абсолютное значений ранговой корреляции, с которым уже можно работать на практике. Также, достоинством этого решения выступает тот факт, что это классификатор, а не регрессия. Но, одновременно, у данного решения есть недостатки. Оно недостаточно гибкое – в качестве входных данных классификатор ожидает переводы именно из тех же источников перевода, на которых он обучается. В данном случае, чтобы получить адекватные результаты, на вход алгоритму должны подаваться четыре перевода: человеческий, машинный статистический, машинный на основе правил и гибридный машинный перевод. Другие решения могут работать с любой структурой входных данных. MatrixNet регрессия показала лучшие результаты по трём ключевым метрикам из четырёх. Также, MatrixNet регрессия не требовательна к отбору факторов и устойчива к переобучению. Все вышеобозначенные тезисы говорят о положительном влиянии использования алгоритма MatrixNet в рамках данной задачи.

Заключение

В данной работе были проведены исследования эффективности существующих факторов и методов машинного обучения, применяемых в задаче оценки качества перевода без использования эталонов перевода. Были предложены новые факторы,с помощью ключевых метрик показано их положительное влияние на качество. Был предложен метод машинного обучения, который на данной задаче показал отличный прирост качества классификатора. Поставленная задача обойти Baseline решение по ключевым метрикам была успешно достигнута.

Основная сложность в данной задаче – качественные факторы, которые будут работать одинакого хорошо для всех видов перевода. Данные, с которыми мы работает в рамках данного направления, очень шумные и сложные для анализа. Именно новые факторы и новые модели, по которым их можно будет подсчитать могут сделать дальнейший скачек в этой задаче.

В дальнейшем, необходимо продолжить искать факторы, которые будут хорошо работать на данной задаче. Также, очень интересно оценить, насколько данные методы подходят для работы с англо-русской языковой парой. Для того, чтобы это оценить, необходимо подготовить обучающий и тестовый наборы, размеченные человеком.

Список литературы

1. *Koehn P.* Statistical Machine Translation //. 2010.
2. *Lin C.-Y., Och F. J.* ORANGE: a Method for Evaluating Automatic Evaluation Metrics for Machine Translation //. 2004.
3. *Bojara O., Buck C., Callison-Burch C., Federmann C., Haddow B., Koehn P., Monz C., Post M., Soricut R., Specia L.* Findings of the 2013 Workshop on Statistical Machine Translation //. 2013.
4. *Callison-Burch C., Koehn P., Monz C., Post M., Soricut R., Specia L.* Findings of the 2012 Workshop on Statistical Machine Translation //. 2012.
5. *Drucker H., Burges C. J. C., Kaufman L., Smola A. J., Vapnik V. N.* Support Vector Regression Machines //. 1997.
6. *J F.* Greedy Function Approximation: A Gradient Boosting Machine //. 1999.
7. Matrixnet. Technical report. 2010. http://www.ashmanov.com/arc/searchconf2010/08gulin-searchconf2010.ppt.
8. TreeNet. http://www.salford-systems.com/products/treenet.
9. *Andrey G., Pavel K., Denis R., Ilya S.* Яндекс на РОМИП'2009. Оптимизация алгоритмов ранжирования методами машинного обучения //. 2009.
10. TreeTagger. http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/.
11. Stanford Parser. http://nlp.stanford.edu/software/lex-parser.shtml.
12. Berkeley Parser. https://code.google.com/p/berkeleyparser/.
13. ACL 2014 NINTH WORKSHOP ON STATISTICAL MACHINE TRANSLATION: Quality Estimation task. http://www.statmt.org/wmt14/quality-estimation-task.html.
14. QuEst - an open source tool for translation quality estimation. http://www.quest.dcs.shef.ac.uk/.
15. QuEst black-box features. http://www.quest.dcs.shef.ac.uk/quest˙files/features˙blackbox.
16. *Shah K., Cohn T., Specia L.* An Investigation on the Effectiveness of Features for Translation Quality Estimation //. 2013.
17. *Arase Y., Zhou M.* Machine Translation Detection from Monolingual Web-Text //. 2013.