Методы пополнения корпусных данных в статистическом машинном переводе

Р. Г. Мифтахова

Башкирский государственный университет
Россия, Республика Башкортостан, 450076 г. Уфа, улица Заки Валиди, 32.

Email: miftahovar@yandex.ru

Для увеличения корпусных данных для систем статистического машинного перевода предложено использование промежуточного корпуса тривиально родственного или родственного языка.

Ключевые слова: статистический машинный перевод, выравнивание, корпус, конкатенация.

Для качественного статистического машинного перевода некоторым языковым парам недостает корпусных данных. Переводоведение сегодня воспринимается как относительно новая наука, находящаяся в процессе становления и претерпевающая определенное развитие. В связи с этим актуальны разработки алгоритмов, которые позволили бы заменить очень трудоемкий и длительный процесс составления новых корпусов. Одним из таких методов является метод «раскручивания». Авторство такой идеи принадлежит С. Баннарду и С. Каллизон – Берчу (2005), которым удалось улучшить качество СМП, основанного на фразах, за счет извлечения парафраз. В качестве исследуемого материала были взяты восемь языков из корпуса Europarl. Обучающи материалом послужили 10 000 пар предложений (испанский – английский). В тестовом предложении обязательно находилось слово, для которого не обнаруживалось перевода в такой незначительной базе. Например, испанское предложение «Es positivo llegar a un acuerdo sobre los procedimientos, pero debemos encargarnos de que este sistema no sea susceptible de ser usado como arma pol´ıtica» было переведено как «It is good reach an agreement on procedures, but we must encargarnos that this system is not susceptible to be usado as political weapon». Слова encargarnos и usado остались непереведенными, так как не были выровнены. Стратегия «раскручивания», которая применялась для непереведенных сегментов, заключалась в подстановке парафраз для этих сегментов, и последующем переводе этих парафраз. Ниже представлены примеры парафраз и их переводы.

Encargarnos – to ensure, take care, ensure that

Garantizar – guarantee, ensure, guaranteed, assure, provided

Velar – ensure, ensuring, safeguard, making sure

Procurar – ensure that, try to, ensure, endeavour to

Asegurarnos – ensure, secure, make certain

Usado – used

Utilizado – used, use, spent, utilized

Empleado – used, spent, employee

Uso – use, used, usage

Utiliza – used, uses, used, being used

Utilizar – to use, use, used

Отсюда следует, что если извлечь перевод для «Garantizar» можно использовать его и как перевод для «Encargarnos», соответственно и перевод для «utilizado» вместо «usado». [Improved Statistical Machine Translation Using Paraphrases Bannard and Callison-Burch (2005)C.3].

Пример использования параллельного корпуса для извлечения парафраз: [Improved Statistical Machine Translation Using Paraphrases Bannard and Callison-Burch (2005)C.3].

what is more, the relevant cost dynamic is completely under control

im übrigen ist die diesbezügliche kostenentwicklung völlig unter kontrolle

wir sind es den steuerzahlern schuldig die kostenunter kontrolle zu haben

we owe it to the taxpayers to keep the costs in check

Из этого примера видно, что фраза «unter kontrolle» извлекла две парафразы «under control» и «in check».

Для каждой парафразы рассчитывается коэффициент вероятности. Такие парафразы добавляются в качестве переводных сегментов в систему статистического машинного перевода, основанного на фразах. Позже такая модель была дополнена функцией, определяющей минимальное расстояние редактирования, которая помогла избавиться от парафраз с низкой вероятностью. Результат такого обновления оказался довольно внушительным, на значение 1.8 по BLEU отслеживалось абсолютное улучшение качество перевода.

Метод «раскручивания» пополняет только лексическую базу языка-источника, тогда как, предложенный автором, метод конкатенации обогащает и язык источника и язык цели. Но у метода «раскручивания» есть и преимущество – он может работать с тремя или более независимыми парами языковых данных -корпусов, а не анализирует только родственные языки. К числу положительных характеристик метода пополнения базы через родственные языки можно отнести такой фактор, как контекст. Метод «раскручивания» его не дифференцирует. За счет использования дополнительного родственного языка в качестве источника перевода достигается коэффициент улучшения качества машинного перевода по алгоритму BLEU, сравнимый с методом «раскручи-

вания» при использовании шести языков. Метод применения родственного языка предполагает использование только одного дополнительного параллельного корпуса, тогда как для метода «раскручивания» требуется два дополнительных корпуса. Более того, оба метода никак не противопоставляются друг другу и могут использоваться в комбинации.

В целом, метод «раскручивания» может быть широко использован в СМП. Он может быть применен для распознавания акронимов. Согласно данным internetsalngnet.com в английском языке в настоящий момент насчитывается 5090 широко используемых акронимов. В таком случае раскручивание может выполняться на уровне предложений или на уровне фраз в несколько этапов, когда белорусские предложения сначала переводятся на русский, затем русские предложения – на английский;

Важно отметить тот факт, что лингвисты Н. Хабаш и Дж. Хью, проведя эксперимент, обнаружили, что при переводе с арабского на китайский метод «раскручивания» показал лучшие результаты, чем непосредственное использование параллельных корпусов для этих языков. А японским разработчикам удалось добиться более точного выравнивания на уровне слов, используя технику «раскручивания».

В целом, модель раскручивания направлена на создание системы машинного перевода для новой языковой пары A-B при наличии двуязычных корпусов A – C и C – B, например с белорусского на русский посредством корпуса английский – русский. Данное же исследование направлено на создание системы, улучшающей перевод не с A на B, а с A на C, т. е. с белорусского на английский. Исследование проводилось на предположении, что корпус A – C недостающий, тогда как корпус C – B достаточный, причем языки A и B близко или тривиально родственные.

Данное исследование было мотивировано следующими факторами. Существующие корпуса для белорусского языка нельзя отнести к достаточным. 5 300 000 неструктурированных слов без морфосинтаксической аннотации представлено корпусом Intercorp (для сравнения в этом же корпусе чешский язык представлен в объеме 174 000 000 слов, и русский – более 13 000 000 слов.) Корпусные данные белорусского языка представлены также в русском национальном корпусе, более в 6 839 000 слов. На время написания данной работы белорусский национальный корпус не создан. В своей статье (2011) «Об основных задачах создания параллельного русско-белорусского корпуса» А. В. Зубов дал описание возможности создания такого проекта.

Для русского языка эти данные более оптимистичны. Национальный корпус русского языка ruscorpora.ru представлен общим объемом 500 000 000 слов. Он характеризуется сбалансированным составом текстов, содержит различные типы устных и письменных текстов с разметкой. Параллельный корпус данного источника содержит более 54 000 000 слов, которые представлены в основном, акцентологическом, мультимедийном, поэтическом, синтаксическом, устном, газетном и диалектном подкорпусах. Кроме того, существует множество других, более мелких русских одноязычных и параллель-

ных англо-русских корпусов, среди них ufal.mff.cuni.cz. Эти сведения дают основание полагать, что корпус белорусского языка недостаточен для качественного статистического машинного перевода, и русский язык (русский корпус) может служить дополнительным инструментом для пополнения белорусско-английского корпуса..

Параллельные тексты для одного языка можно использовать с целью улучшения статистического машинного перевода для близко или тривиально родственных языков. Для соответствия правилам орфографии целевого языка возможно также применение алгоритмов транслитерации однокоренных слов.

Целью данного исследования является определение степени вероятности улучшения качества СМП, основанного на фразах, с белорусского на английский язык через использование двуязычного русско – английского корпуса, с предположением, что белорусско-английский корпус недостаточный. Работа состояла из трех основных этапов: 1. конк×1- была произведена простая конкатенация двух корпусов: оригинального белорусско-английского и дополнительного русско-английского, 2. конк×k – произведена конкатенация нескольких копий оригинального корпуса и одной копии дополнительного, так, чтобы в среднем объемы оригинальных и дополнительных данных были одинаковы, 2. конк×k:вырав – произведена конкатенация нескольких копий оригинального корпуса и одной копии дополнительного, сформирован на его основе новый обучающий корпус. Из полученного корпуса сгенерировано выравнивание на уровне слов, затем все повторные копии выровненных слов удалены и оставлена только одна копия для каждого слова оригинала.

Конкатенация параллельных корпусов

Какие преимущества можно извлечь из конкатенации двух параллельных корпусов (белорусско-английского и русско-английского) в один большой параллельный корпус?

Во-первых, такой метод может улучшить выравнивание на уровне слов для предложений белорусско-английского языка, особенно это касается так называемых «редких» слов. Дополнительные предложения из русского корпуса добавляют новые контексты для таких слов, что потенциально улучшает выравнивание для них, а, в результате, и выравнивание на уровне фраз.

Такие «редкие» слова упоминаются в работах Д. Пьетра и др. Он называет их «мусоросборником» при применении моделей выравнивания IBM. Проблема заключается в том, что “редкие”слова могут выравниваться с множеством слов целевого языка. Эту проблему детально рассмотрели К Ганчев и Б. Таскар в 2010 году.

Еще одним преимуществом конкатенации является добавление переводных вариантов, новых неустойчивых фраз для языка-источника, что увеличивает лексическую составляющую, обеспечивает больший выбор и приводит к уменьшению количества неизвестных слов и выдаче более естественного перевода.

Однако простая конкатенация может быть довольно проблематичной. Если конкатенировать небольшой белорусско-английский корпус с объемным русско-английским, последний будет доминировать при выравнивании слов и извлечении фраз, что сильно повлияет на значения вероятности при переводе и приведет к плохой производительности СМП. Избежать такого исхода можно, если добавлять русско-английский корпус только в таком объеме, который бы не превышал оригинальный белорусско-английский корпус. И, так как корпуса объединяются механически, различить фразы, исходящие из белорусско-английского корпуса от фраз, исходящих из русско-английского корпуса невозможно. Хотя было бы предпочтительно, чтобы в первую очередь выбирались фразы из белорусско-английского корпуса, потому что источником перевода в нашем случае является белорусский язык. Такой подход частично решил бы проблему с так называемыми «ложными друзьями» переводчика. В белорусском языке таких слов насчитывается около 100. Например, белорусское слово «лік» означает «число», урадлівы– урожайный, плот– забор, твар– лицо, трус– кролик, лістопад – ноябрь, качка– утка, блага– плохо, бялізна– белье. И если СМП будет брать эти слова из русско-английского корпуса, их перевод явно окажется неадекватным.

Исходя из этих обоснований данное исследование проводилось по следующим направлениям:

- были объеденены два корпуса (белорусско – английский и русско-английский) в один корпус для обучения СМП, основанного на фразах.

- были объеденены 2 копии белорусско – английского корпуса и 1 корпус русско-английский, такая пропорция обеспечивала приблизительно идентичные объемы этих корпусов.

Еще одним преимуществом использования близко или тривиально родственных языков при недостаточных корпусных данных является построение таблиц с выровненными фразами. Таблицы с выровненными фразами из белорусско-английского и русско-английского корпуса могут также быть использованы для улучшения белорусско-английского СМП. Их можно использовать в качестве альтернативных способов декодирования [П. Коен, 2007] или объединить и присвоить фразам дополнительные метки, что позволит идентифицировать к какому корпусу они принадлежат. В нашем эксперименте при конкатенации приоритеты отдавались корпусу белорусский-анлийский “БА”, т.е. этот корпус был взят как основа, и фразы, которые в нем отсутствовали, добавлялись из корпуса русский – английский “РА”. Для каждой добавленной фразы сохранялись значения вероятностей перевода (значения условной вероятности). Полученная комбинированная таблица была маркирована тремя дополнительными метками A1, A2, A3. A1=1, если фраза из корпуса “БА”, в противном случае A1=0.5; A2=1 если фраза из корпуса “РА”, в противном случае A2=0.5; A3=1, если фраза присутствовала и в том и в другом корпусе, в противном случае A3=0.5. В результате

было получено три комбинации меток [1;0.5;0.5], [0.5;1;0.5], [1;1;1]. Последняя метка для фраз, присутствующих в обоих корпусах.

Эксперимент проводился в три этапа. Сначала был использован только параметр A1, затем A1 и A2, и, наконец, A1, A2, A3. С помощью алгоритма MERT [F. Och. Minimum Error Rate Training in Statistical Machine Translation. Information Sciences Institute University of Southern California 4676 Admiralty Way, Suite 1001 Marina del Rey, CA 90292] каждому параметру пары фраз (пяти стандартным параметрам и трем дополнительным) присваивался вес. Этот вес позволял определить оптимальное количество дополнительных параметров A1, A2, A3. Т.е. сначала был применен параметр A1, затем A1 и A2, и, наконец, A1, A2, A3. В результате была выбрана таблица фраз, получившая наибольшее значение BLEU (по методу С. Накова, 2008).

В ходе эксперимента была предпринята попытка получить оптимальные результаты из вышерассмотренных методов, учитывая все их преимущества и недостатки, а именно а) получить улучшенное выравнивание на уровне слов белорусско-английского корпуса путем добавления пар предложений из русско- английского корпуса, б) увеличить лексическую составляющую белорусско-английского корпуса за счет использования дополнительной таблицы фраз из русско- английского корпуса.

Такой подход был аргументирован наличием однокоренных слов в близко или тривиально родственных языках. Однако если классическая лингвистика определяет однокоренные слова, как слова, имеющие общее происхождение, общий корень с его лексическим значением, то компьютерная лингвистика не учитывает происхождение слова. Она определяет однокоренные слова, как слова в разных языках, которые являются взаимными переводами друг друга и имеют схожую орфографию.

Качество машинного перевода для языков с недостаточной базой одноязычных и параллельных данных может быть повышено через использование того преимущества, которое обеспечивает их аналогия с языками, обладающими значительными лингвистическими ресурсами. Таким образом, можно улучшить перевод с белорусского языка с недостаточной базой Б на английский язык со значительной базой Е, используя ограниченный параллельный корпус Б – Е и более содержательный корпус Р-Е (русско-английский), учитывая тот факт, что русский язык очень близок к белорусскому. Лексическое сходство между двумя языками, а также сходство правописания и синтаксиса позволяют улучшить выравнивание (alignment) строк для языка с недостаточной базой, применять дополнительные алгоритмы при переводе, а также, используя метод нормализации, учитывать потенциальные различия в правописании. В результате проведенного исследования оценка машинного перевода с белорусского на английский язык с использованием корпуса русского языка показала абсолютное улучшение качества на 0.51 по алгоритму BLEU, что является значительным показателем.

Применение данного метода "обогащает" исходную базу языка как минимум в 1.6 раза. Эти данные могут варьироваться в разных языковых парах в зависимости от того, насколько схожи родственные языки.

Литература

1. Кипяткова И. С. Применение синтаксического анализа при создании n-граммной модели языка для систем распознавания русской речи / И. С. Кипяткова // Труды 5 междисциплинарного семинара Анализ разговорной русской речи АРЗ–2011, 25–26 августа 2011 г. – СПб., 2011. – С. 13–18.
2. Бабин Д. Н. О перспективах создания системы автоматического распознавания слитной устной русской речи / Д. Н. Бабин, И. Л. Мазуренко, А. Б. Хо-лоденко // Интеллектуальные системы. – 2004. – Т. 8, Вып. 1–4. – С.45–70.
3. Pang B., Lee L. Sentiment Classification using Machine Learning Techniques. Proceedings of the Conference on Empirical Methods in Natural. Language Processing (EMNLP). Philadelphia. 2002. P. 79–86.
4. Бузикашвили Н. Е., Самойлов Д. В., Бродский Л. И., Усков А. В. Задача поиска в неструктурированном тексте и лингвистический анализ. // Интеллектуальные технологии ввода и обработки информации, М., 1998.
5. Морозкина Е. А., Влияние информационных технологий на развитие лингвистических норм. // Вестник Башкирского университета, -2012 №1 -С 163.
6. Морозкина Е. А., Наука о переводе в свете лингвистического учения Вильгельма Фон Гумбольдта // Языки в диалоге культур, Материалы 2 Международной научно-практической конференции, посвященной 100-летию со дня рождения первого ректора БашГуШ. Х. Чанбарисова. -2016. -С 33
7. URL: http://www.rae.ru/monographs/189–5958

Статья рекомендована к печати кафедрой лингводидактики и переводоведения БашГУ

New approaches to resolving the problem of corpora data shortage

R. G. Miftakhova

Bashkir State University
32 Zaki Validi Street, 450074 Ufa, Republic of Bashkortostan, Russia.

Email: miftahovar@yandex.ru

The usage of agnate languages for more accurate alignment in statistical machine translation to resolve the problem of corpora data shortage.

Keywords: statistical machine translation, concatenation, corpora, alignment.