ВЫРАВНИВАНИЕ НЕРАЗМЕЧЕННОГО КОРПУСА ПАРАЛЛЕЛЬНЫХ ТЕКСТОВ

ALIGNMENT OF UN-ANNOTATED PARALLEL CORPORA

Потемкин С.Б. *(potemkin@philol.msu.ru),* ***Кедрова Г.Е.*** *(kedr@philol.msu.ru)*
Московский государственный университет им. М.В. Ломоносова

Предлагается два связанных между собой алгоритма выравнивания параллельных текстов на уровне предложений и на пословно-пооборотном уровне. Метод основан на сопоставлении редко встречающихся слов в обоих текстах с последующим применением динамического программирования.

Введение

Выравнивание параллельных текстов, то есть автоматическое сопоставление предложений или слов в одном тексте их эквивалентам в переводе, является очень важным этапом предварительной обработки для многих приложений, включая, помимо прочего, машинный перевод (Brown и др., 1993), информационный поиск по текстам на различных языках (Hiemstra, 1996), составление словарей (Smadja и др., 1996) и получение данных для обработки естественного языка (Kuhn, 2004), создание корпуса параллельных текстов (Гарабик, Захаров, 2006).

Наиболее широко применяемые статистические методы, которые не требуют развитой словарной базы и могут использоваться для редких языков, часто дают ошибочные результаты выравнивания, требуя в последующем дорогостоящей ручной проверки и исправления. Использование двуязычных словарей для выравнивания текстов менее распространено, и применялось в основном для специализированных текстов, в то время как для художественных текстов, в которых гораздо чаще встречаются неоднозначные соответствия между предложениями, примеры применения метода приводятся редко.

В отличие от известных методик, в качестве элементов матрицы смежности рассматриваются не отдельные сопоставленные слова, а интервалы от одного пробела в предложении до следующего. Это дает возможность сопоставлять словарные словосочетания из одного предложения – слову или словосочетанию из другого.

1. Предшествующие работы

Выравнивание по предложениям

Методы выравнивания по предложениям относятся, грубо говоря, к трем категориям:

- На основании длины (Gale и Church, 1993), в предположении, что длина предложения в оригинале и в переводе примерно совпадают.
- На основании двуязычной лексической информации, например полученной из корпуса (Kay и Ruoscheisen, 1993; Fung и Church, 1994; Fung и McKeown, 1994).
- Алгоритмы с привлечением опорных меток выравнивают предложения на основании информации, содержащейся в размеченном корпусе или по орфографическому сходству (Simard и др., 1992).

Методы на основании длины очень чувствительны к пропускам в том смысле, что отдельный пропуск может приводить к неправильному последующему выравниванию от точки пропуска до конца корпуса.

Для вычисления сходства двух структурных единиц текстов вводится некоторая мера близости, например количество переводных эквивалентов, имеющихся в словаре. Полученный вес нормализуется на длину текста, чтобы величины для разных единиц текста были сопоставимы. (А.Ф. Гельбух и др., 2006).

Для оптимального решения задачи выравнивания применяется метод динамического программирования. Однако вычисление всей матрицы сопоставления для достаточно больших текстов не представляется возможным вследствие больших временных затрат. Поэтому вместо решения задачи на всей матрице предлагается решать её на некоторой окрестности диагонали (Липатов, Мальцев, 2006).

Выравнивание по словам

Выравнивание на уровне ниже уровня предложений обычно выполняется с использованием статистических моделей для машинного перевода (Brown и др., 1993; Hiemstra, 1996; Vogel и др., 1999) где любое слово целевого языка считается возможным переводом для каждого слова исходного языка. Для каждой пары слов исходного и целевого текстов выписывается число сегментов, которые (a) содержат оба слова, (b) содержат слово исходного языка не содержат слово целевого языка (c) слово целевого языка, но не слово исходного языка и (d) ни то ни другое слово (Ribeiro и др., 2000).

Найденные таким образом наиболее вероятные пары принимаются в качестве переводных эквивалентов. Такой подход имеет ряд недостатков, связанных с большим количеством редких слов, обычным для любого корпуса, различиями в порядке слов в языках, между которыми производится выравнивание, и наличие словосочетаний. Аналогичным образом построен алгоритм (Липатов, Мальцев, 2006), но с использованием имеющегося англо-русского словаря. Наилучшая пара эквивалентов отыскивается в параллельных предложениях локально, без учета оставшейся части предложения. При этом возможно сопоставление одному слову русского языка нескольких эквивалентов английского, и наоборот.

Другая проблема касается различий в порядке слов между исходным и целевым языком. В предположении локальности инверсии в порядке, который может быть оформлен в пределах некоторого окна, возможна обработка этого явления (Потемкин, Кедрова, 2007).

Подводя итог, отметим, что выравнивание, как на уровне предложений, так и на более низком уровне (слова, словосочетания), нуждается в совершенствовании: существующие модели не позволяют выравнивать тексты с пропущенными или несовпадающими предложениями, редкие слова и словосочетания в пределах предложений. Синтаксические различия между исходными и целевыми языками также представляют затруднения для большинства стратегий выравнивания.

2. Предлагаемый метод выравнивания по предложениям

Основной проблемой при автоматическом выравнивании текста на уровне предложений является появление ложных пар предложений, полученных при работе алгоритма, но не являющихся переводными эквивалентами. При разработке предлагаемого метода мы старались свести к минимуму такие явления.

Алгоритм выравнивания разработан в предположении, что (a) порядок предложений в русском и английском текстах совпадает (b) в параллельных текстах нет значительных (более 500 слов) пропусков (c) длина текстов не слишком большая – рассказ или глава романа, около 64. Последние ограничения непринципиальны и связаны в основном с временем работы алгоритма. Метод основан на использовании обширного англо-русского словаря (Кедрова, Потемкин, 2005), по которому выполняется поиск переводных эквивалентов из анализируемых текстов. В отличие от статистических методов и методов, основанных на мере близости, предлагается рассматривать только низкочастотные слова, а именно слова, встречающиеся только 1 раз в каждом тексте (*hapax legomena*). Вначале для каждого такого слова (русского) текста находим переводной эквивалент в (английском) тексте, который также имеет частотность 1. Если для этого русского слова нашлось несколько эквивалентов, все они исключаются из рассмотрения. Если, далее, найденные эквиваленты связывают предложения с нарушением их порядка в тексте, они также исключаются. В результате такой ограничительной стратегии получаем набор уникальных пар эквивалентов в двух текстах. Такие пары образуют первичную структуру опорных точек или якорей, связывающих те предложения текстов, к которым они относятся. На этом этапе мы не говорим об эквивалентности найденных пар предложений, можно только утверждать, что такие пары предложений имеют непустое пересечение. Затем исходные тексты разбиваются на отрезки, ограниченные найденными парами предложений. Эти отрезки рассматриваются как новые параллельные тексты и процедура расстановки опорных точек повторяется. Итерации продолжаются, пока появляются новые якоря. На практике число итераций в рассмотренных текстах не превышает 6.

После определения опорных точек производится поиск критического пути методом динамического программирования. Для этого рассматриваются отрезки русского и английского текста между опорными точками. Для каждого слова отрезка русского текста отыскивается словарный эквивалент в соответствующем отрезке английского текста. Число таких эквивалентов подсчитывается для каждой пары предложений. Полученные таким образом меры сходства нормируются на единицу по длине предложений и записываются в матрицу смежности. К элементам матрицы, соответствующим опорным точкам к значению меры сходства прибавлено большое число (10000), чтобы критический путь заведомо прошел через эти точки. Далее поиск критического пути выполнялся стандартными методами динамического программирования («поиск Витерби»).

3. Эксперименты по выравниванию на уровне предложений

Для оценки работы этой части алгоритма было произведено выравнивание нескольких текстов произведений русской классики (Н.В. Гоголь, Ф.М. Достоевский, А. П. Чехов) На рис. иллюстрируются результаты эксперимента, проведенного для рассказа А.П. Чехова «Анна на шее» и его перевода на английский язык. Тексты не подвергались никакой предварительной обработке или разметке. В качестве границы предложения приняты точка, восклицательный знак, вопросительный знак, многоточие.

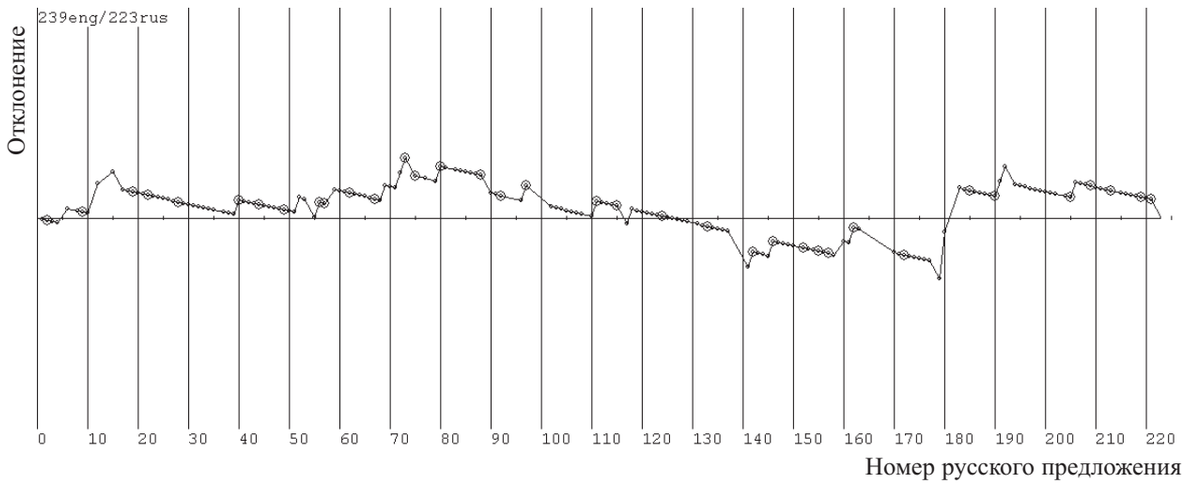

Рис. 1 Диаграмма выравнивания (Гоголь, «Шинель)
Показаны точки критического пути. Опорные точки выделены кружками.

Горизонтальная линия представляет среднее = отношение числа английских предложений к числу русских предложений. По оси Y показано отклонение номера английского предложения от среднего.

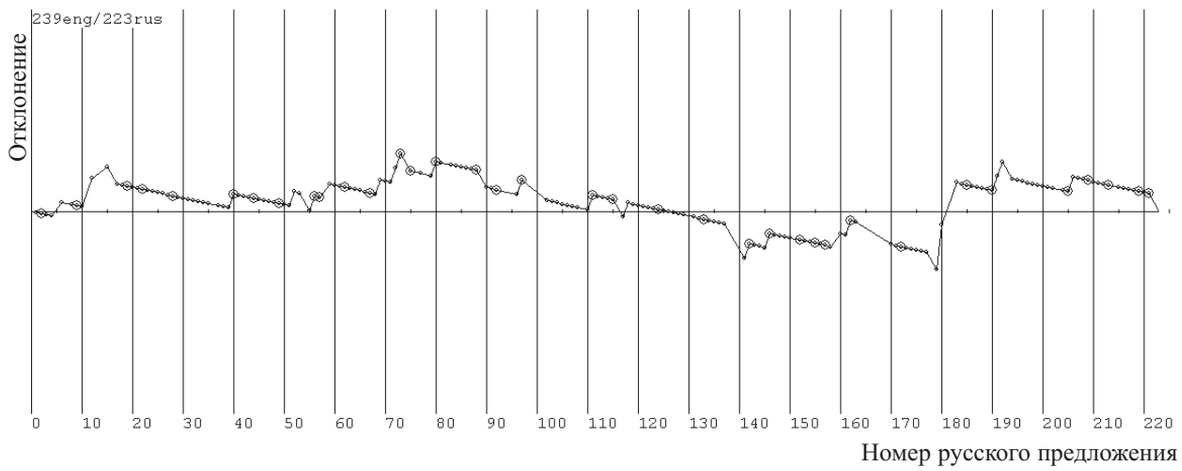

Рис. 2 Диаграмма выравнивания (Чехов, «Анна на шее»)

При выделении предложений учитываются некоторые распространенные сокращения, содержащие точку (Mrs. Mr. Ms. Prof. Dr. Gen. Rep. Sen. St. etc. i.e. e.g. et al.; т.д. т.к. пр. Св. и другие). Эти сокращения распознаются, и точка в этом случае не считается концом предложения. Не учитываются символы перевода строки, заглавные буквы, двоеточия, кавычки. Отказ от учета перевода строки связан с тем, что многие тексты в электронном виде получены в результате сканирования и перевод строки в этом случае может не совпадать с авторским делением на абзацы. Кроме того, в переводе деление на абзацы, также как выделение прямой речи и расстановка других знаков препинания, как правило, отличается от оригинала.

В результате работы алгоритма получено 182 пары предложений (78% текстов). Из них 165 предложений (90.5%) являются полным и точным переводом, 16 предложений (9%) являются частью перевода оригинала (или

наоборот) и 1 предложение (0.5%) сопоставлено переводу ошибочно. Аналогичные соотношения сохраняются для других текстов («Шинель» Гоголя, главы из романа «Преступление и наказание» Достоевского).

Типичный пример сопоставления части предложения целому:

*«**Статский советник...** принят у его сиятельства...« или: «Со средствами ...*

=

visits at His Excellency's « ; or , «A man of means…

(несопоставленная часть выделена жирным шрифтом)

Расхождение вызвано различной расстановкой знаков препинания в русском и английском тексте, а также элиминацией части текста при переводе.

Обработка таких несовпадений выполняется второй частью алгоритма.

Чаще всего бывает, что 2 последовательных предложения одного текста (i, i+1) сопоставлены 2 предложениям другого текста, не являющимся последовательными (j, j+2). (Рис. 3)

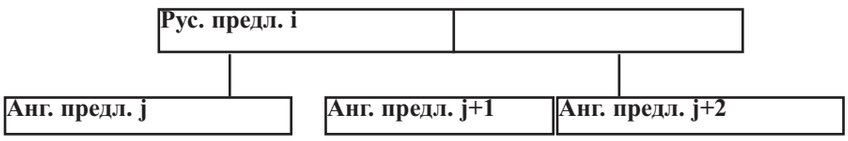

Рис. 3 Объединение 2 предложений при выравнивании

Тогда предполагается, что предложение j+1 является переводом либо части предложения i, либо i+1. Для определения, к какому именно предложению относится j+1 применяется процедура, аналогичная первоначальной разметке, то есть отыскиваются уникальные слова в предложении i, i+1, j+1, их переводы и определяется соответствие, после чего предложение j+1 сливается либо с j, либо с j+2.

Результаты, полученные на основе нашего алгоритма сравнивались с аналогичными (ABBYY, 2007) на тексте «Шинель» Гоголя (433 пары предложений). Доля предложений, правильно сопоставленных алгоритмом ABBYY составила 71%.

4. Фрагментация параллельных предложений

В качестве координатных отсчетов пространства билингвы будем принимать не слова как таковые, а разделители (пробелы) между соседними словами (Потемкин, Кедрова, 2005). При таком подходе отображение слова исходного предложения на слово целевого предложения представляет собой отрезок с координатами начала и конца слова SS по x и начала и конец слова TS по y. Теперь возможно ставить в соответствие не только однословные эквиваленты, но также эквиваленты типа словосочетаний. На рис. 1 показано отображение SS на TS, выполненное с учетом встретившегося в словаре словосочетания (*вдруг == all at once*) Коллизия возникает, когда некоторые отображающие отрезки перекрываются по горизонтали или по вертикали (т.е. отображение не однозначное). Например, слово исходного текста к отображается на слова целевого текста *at, to, with*. Чаще всего в коллизии участвуют служебные слова, а также знаки препинания.

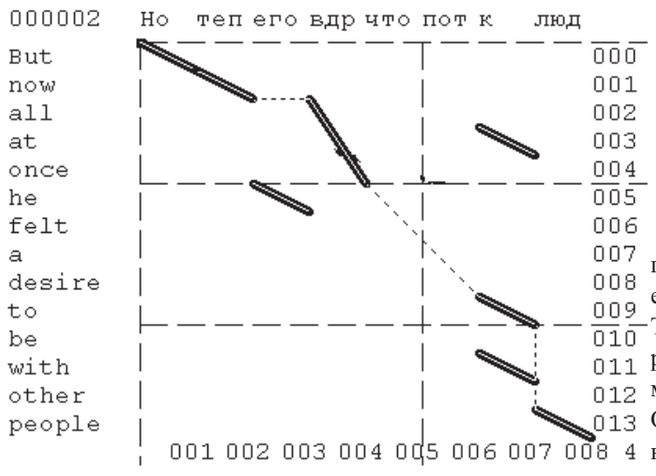

Рис. 4 Отображение SS «Но теперь его вдруг что-то потянуло к людям» на TS «But now all at once he felt a desire to be with other people» (Ф.М.Достоевский) Пунктиром показан критический путь.

Построив пословное отображение, можно переходить непосредственно к фрагментации, то есть к отображению интервалов SS на интервалы TS, которые лежат между уже построенными отображающими отрезками. Слова двух предложений можно сопоставлять в разной последовательности. Одно и то же предложение может быть переведено как с прямым, так и с обратным порядком слов и оба перевода будут правильными. Более общий слу-

чай – когда одни группы слов переведены в прямом направлении, другие – в инверсном и сами эти группы сопоставляются хаотически.

Данную задачу можно свести к задаче нахождения максимального паросочетания на двудольном графе. Такая задача решается методом нахождения максимального потока на сети или методом, разработанным специально для сетей с единственным истоком и единственным стоком (Диниц, 1970). К сожалению, эксперименты по применению этого метода в настоящее время не привели к положительным результатам по фрагментации предложений с несовпадающим порядком слов в SS и TS.

В случае, если мы рассматриваем только монотонные отображения (т.е. считаем порядок слов исходного и целевого предложения по большей части совпадающим), задача попадает в класс более простых задач динамического программирования.

Как правило, исходное предложение и его перевод, даже имеющие в целом совпадающий порядок слов, содержат фрагменты с инверсией, напр. *{изредка только; only occasionally}*. Такую частичную инверсию желательно включить в критический путь, но приведенный алгоритм этого не допускает. Предлагается формировать фиктивное отображение для инверсных фрагментов заданной длины (напр., не больше 3 слов), которые включаются в общий набор отображающих отрезков и участвуют в алгоритме поиска критического пути.

Вернемся теперь к предложению рис. 1.

Критический путь разбивает исходную пару предложений на следующие фрагменты:

1. Но теперь == But now
2. теперь его вдруг == now all at once
3. вдруг что-то потянуло к == all at once he felt a desire to be with
4. к людям == with other people

Фрагмент 1 не вызывает возражений. Фрагмент 2 SS содержит местоимение его, которого нет во фрагменте TS. Наоборот, фрагмент 3 TS содержит местоимение he, которое отсутствует во фрагменте SS. Если мы объединим фрагменты 2 и 3, результат будет более осмысленным. При решении вопроса об объединении фрагментов мы придерживаемся двух критериев: a) отношение длин фрагментов SS и TS не должно сильно отличаться от 1 и b) переводы всех слов, содержащихся в SS должны содержаться в TS и наоборот. Фрагмент 4 является конечным или хвостовым, и уже не может быть объединен со следующим и его оценка по любому из критериев a) или b) не производится. Поэтому доверие к хвостовому фрагменту заведомо ниже, чем к начальным.

Эксперименты проводились на корпусе параллельных литературных текстов и текстов юридического содержания, на котором метод дает гораздо лучшие результаты. В развитие данного подхода будет составлен автоматический словарь фрагментов для использования в системе автоматического перевода, основанного на примерах (EBMT) (Потемкин, Кедрова, 2007).

Список литературы

1. Brown P.F., Della Pietra S.A., Della Pietra V.J., Mercer R.L., 1993. The mathematics of machine translation: Parameter estimation. // Computational Linguistics, 19(2):263–311.

2. Dinic E.A. An algorithm for the solution of the max-flow problem with the polynomial estimation. Dokl. Akad. Nauk SSSR 194 (1970), no.4 (in Russian); English transl.: Soviet Math. Dokl. 11 (1970), 1277-1280

3. Fung P., McKeown K., 1994. Aligning noisy parallel corpora across language groups: word pair feature matching by dynamic time warping. // Proceedings of the First Conference of the Association for Machine Translation in the Americas (AMTA-94), pages 81–88, Columbia, Maryland, USA.

4. Gale W.A., Church K.W., 1993. A program for aligning sentences in bilingual corpora. // Computational Linguistics, 19(1):75–102.

5. Kay M., Roscheisen M., 1993. Texttranslation alignment. // Computational Linguistics, 19(1):121–142.

6. Kuhn J., 2004. Exploiting parallel corpora for monolingual grammar induction – a pilot study. // Workshop proceedings of the 4th International Conference on Language Resources and Evaluation (LREC), pages 54–57, Lisbon, Portugal. LREC Workshop: The Amazing Utility of Parallel and Comparable Corpora.

7. Ribeiro A., Lopes G., Mexia J., 2000 A Self-Learning Method of Parallel Texts Alignment // J.S. White (Ed.): AMTA 2000, LNAI 1934, pp. 30−39, 2000.Springer-Verlag Berlin Heidelberg 2000

8. Schrader B., 2006 ATLAS – a new text alignment architecture // Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions, pages 715–722, Sydney, July 2006. Association for Computational Linguistics

9. Simard M., Foster G.F., Isabelle P., 1992. Using cognates to align sentences in bilingual corpora. // Proceedings of the Fourth International conference on theoretical and methodological issues in Machine translation, pages 67–81, Montreal, Canada.

10. Vogel S., Ney H., Tillmann C., 1999. HMM-based word alignment in statistical translation. // Proceedings of the International Conference on Computational Linguistics, pages 836–841, Copenhagen, Denmark.

11. ABBYY, 2007 Выравниватель предложений текстов на русском и английском языке, http://webaligner.abbyy.com/

12 Гарабик Радован, Захаров Виктор Параллельный русско-словацкий корпус // Труды международной конференции Корпусная лингвистика – 2006. Sankt-Petersburg: St. Petersburg University Press 2006

13. Гельбух А.Ф., Сидоров Г.О., Вера-Феликс А., 2006. Словари в задачах автоматической обработки пар переводных текстов // Труды межд. конференции Диалог-2006, М., стр. 110-114.

14. Кедрова Г.Е., Потемкин С.Б., 2004 Семантическое разделение омонимов с использованием двуязычного словаря и словаря синонимов, // II Международный конгресс исследователей русского языка «Русский язык: исторические судьбы и современность», Доклады, М 2004

15. Крылов С.А., Старостин С.А., Актуальные задачи морфологического анализа и синтеза в интегрированной информационной среде STARLING // Труды международной конференции Диалог'2003, М.2003

16. Липатов А.А., Мальцев А.А., 2006 Методы автоматизации построения и пополнения двуязычных словарей с использованием корпуса параллельных текстов // Труды международной конференции Диалог'2006, М.2003

17. Потемкин С.Б., Кедрова Г.Е., 2005 Автоматическая оценка качества машинного перевода на основе семантической метрики // Труды II Международной научно-практической конференции, посвященной Европейскому Дню языков, Луганск 2005.

18. Потемкин С.Б., Кедрова Г.Е., 2007 Использование корпуса параллельных текстов для пополнения специализированного двуязычного словаря // III Международный конгресс исследователей русского языка «Русский язык: исторические судьбы и современность», Доклады, М 2007