СЛОВАРИ В ЗАДАЧАХ АВТОМАТИЧЕСКОЙ ОБРАБОТКИ ПАР ПЕРЕВОДНЫХ ТЕКТОВ[1]

DICTIONARIES IN TASKS OF AUTOMATIC PROCESSING OF PAIRS OF TRANSLATED TEXTS

А.Ф. Гельбух *(www.Gelbukh.com),*

Г.О. Сидоров *(sidorov@cic.ipn.mx),*

А. Вера-Феликс

Лаборатория естественного языка и обработки текста,
Центр Компьютерных Исследований (CIC),
Национальный Политехнический Институт (IPN),
г. Мехико, Мексика

Рассматривается проблема автоматической обработки пар переводных текстов, также часто называемых параллельными текстами. Одна их наиболее важных задач состоит в установлении соответствий между текстами на уровне абзацев, предложений и отдельных слов. На практике это соответствие очень часто не является взаимнооднозначным: одному абзацу могут соответствовать несколько, какие-то слова при переводе просто опускаются или заменяются очень дальними синонимами или фразеологическими оборотами, которые могут быть абсолютно различными в разных языках, и т.п. Обсуждается применение метода, основанного на лексической информации, для автоматического установления таких соответствий для корпуса художественных текстов. Метод основан на том, что большинство слов одного текста все-таки имеет в другом тексте пары прямой перевод, даваемый для этого слова в двуязычном словаре. Приводятся данные экспериментов по выравниванию англо-испанских пар текстов на уровне абзацев, которые показывают, что подобные методы применимы для художественных текстов.

Введение

Лингвистические данные, содержащие разметку любой природы, являются исключительно ценным ресурсом, так как на разработку и создание таких ресурсов затрачивается много ручного труда квалифицированных лингвистов. Возникает естественный вопрос о возможности избежать этих затрат, получая, получая размеченные лингвистические данные автоматически. Этот вопрос является одним из центральных на нынешнем этапе исследований по компьютерной лингвистике. Широко используются для этой цели методы автоматического обучения. Бум в этой области связан также с развитием Интернета, который сделал доступными для автоматического анализа воистину гигантские массивы текстов на самых разных языках.

Чаще всего объектом анализа является текст на одном конкретном языке. Однако дополнительным важным источником получения ценной лингвистической информации являются пары переводных текстов, или, как их еще называют «параллельные тексты». Процесс установления соответствий между парами таких текстов на разных уровнях называют выравниванием. Выравнивание возможно на уровне абзацев, предложений и отдельных слов: устанавливается соответствие между тем, какие абзацы (предложения, слова) одного текста соответствуют каким абзацам (предложениям, словам) другого текста. Такие соответствия не всегда просто установить, потому что одна структурная единица может соответствовать нескольким или быть опущена. Особенно часто отсутствие однозначного соответствия между единицами разных уровней в парах текстов характерно для переводов текстов художественной литературы: в этом случае переводчик старается не только подойти более «творчески» к переводу текста, но и по мере возможности адаптировать его для восприятия аудитории своего культурного ареала.

Имея выравненные параллельные тексты, можно извекать из них много различной полезной лингвистической информации (о смыслах слов, об идиоматических выражениях, о синтаксических констукциях,

[1] Работа выполнена при частичной поддержке правительства Мексики (КОНАЦиТ, СНИ) и Национального Политехнического Института (Мексика). Work done under partial support of Mexican Government (CONACyT, SNI) and National Polytechnic Institute (SIP, COFAA, PIFI), Mexico.

и пр.), воспользовавшись тем, что потому что разные языки представляют информацию не тождественно. В этом смысле разметка параллельных текстов дает эффект, эквивалентный ручной разметке самых разных лингвистических явлений. Таким образом, выравненные пары текстов являются очень полезным лингвистическим ресурсом.

Следовательно, возникает вопрос о том, как проводить разметку соответствий в параллельных текстах. Проводить ее вручную достаточно дорого и трудоемко. Поэтому исследователями разрабатываются разнообразные методы автоматического выравнивания текстов. Как обычно бывает в задачах автоматической обработки текстов, эти методы можно разделить на две группы: методы, основанные на привлечении дополнительных лингвистических знаний, например, (Kay and Roscheisen, 1993, Kit *et al.*, 2004), и статистические методы, например, (Brown *et al.*, 1991; Gale and Church, 1991; McEnery and Oakes, 1996; Mikhailov, 2001). Заметим, что разница состоит в основном в типе используемых данных.

Традиционно для выравнивания параллельных текстов используются статистические методы. Они основаны на том, что известно примерное соответствие длины текстов на одном языке и длины текстов на другом языке, измеренное в словах или в символах. Следовательно, исходя из длины текстов на одном языке, можно вычислить, где должны начинаться соответствующие фрагменты текста на другом языке. Остается только выбрать наиболее подходящее соответствие. Методы этого класса имеют очень высокую скорость работы, потому что только подсчитывают количество слов или количество символов.

Однако у этих методов есть и свои недостатки. Как только появляются случаи значительно более короткого или более длинного перевода, то есть, когда несколько кандидатов претендуют на роль возможного соответствия, алгоритмы, основанные на этих методах, не могут принять решение, потому что информация, используемая ими, очень ограничена.

Другой недостаток этих методов состоит в том, что они по самой своей природе не могут применяться для выравнивания текстов на уровне слов. Действительно, абзацы и предложения обычно являются достаточно крупными единицами, и их длины соответствуют средним величинам для данного языка. Слова же имеют очень небольшую длину, которая к тому же сравнительно случайна в каждом языке. Кроме того, соответствия в порядке слов, в отличии от последовательности абзацев и предложений, являются достаточно свободными, даже в языках с фиксированным порядком слов. Не говоря уже о том, что при переводе именно слова зачастую опускаются или, наоборот, переводятся целым оборотом.

Идея использования лексической (словарной) информации для выравнивания текстов не нова, см., например, (Kay and Roscheisen, 1993; Meyers *et al*, 1998; Chen, 1993; Langlais *et al*, 1998). Однако эти методы не пользуются такой популярностью, как статитситческие методы. Вероятно, одной их причин этого является факт труднодоступности словарей для обработки и трудности с подключением систем автоматического морфологического анализа для взаимной идентификации слов в словарях. Из последних работ на эту тему упомянем (Kit *et al*., 2004). Необходимо заметить, что практически все эксперименты, поставленные для выравнивания параллельных текстов с использованием словарных данных, проводились для достаточно специализированных текстов, вроде текстов канадского или европейского парламентов, юридических текстов, архивов помощи Микрософт, и т. п. Встает вопрос, будет ли такой метод применим для художественных текстов, в которых изначально нет такой параллельности, как в специализированных текстах.

В данной статье мы описываем применение метода выравнивания пар переводных текстов, основанного на дополнительной лингвистической информации, а именно, на двуязычных словарях, к художественным текстам. Сначала мы приводим описание этого метода, затем результаты экспериментов на уровне абзацев, и после этого кратко описываем англо-испанский корпус, который мы построили для использования в экспериментах и последующей работы в направлении исследования параллельных текстов.

Предлагаемый метод выравнивания текстов

Для выравнивания пар переводных текстов мы применяем метод, основанный на вычислении сходства структурных единиц (абзацев, предложений) на основе двуязычных переводных словарей.

Основная идея метода состоит в том, что если слово встретилось в тексте на одном языке, то ожидается, что один из его переводов, приведенных в двуязычном словаре, будет присутствовать в тексте на другом языке. Это не так в случае идиом или вольных переводов, но в большинстве случаев это так. Забегая вперед, можно сказать, что проблемы возникают в указанных случаях при последовательности очень коротких абзацев или предложений, когда не хватает материала для выравнивания.

Заметим, что к выравниванию на уровне слов этот метод применим только после некоторых изменений алгоритма, связанных с необходимостью учета синтаксических структур предложений, потому что только знать соответствия слов без их конкретного местонахождения недостаточно.

Для вычисления сходства двух структурных единиц текстов надо ввести какую-нибудь меру. Как уже понятно из идеи метода, в качестве такой меры мы используем коэффициент сходства, основанный на количестве «общих» (в смысле их соответствия в двуязычном словаре) слов текста и его перевода, вычисленный с учетом всех возможных переводов из словаря. Полученный вес нормализуется на длину текста, чтобы полученные величины для разных единиц текста были сопоставимы.

Чтобы сравнивать тексты, а также чтобы искать переводы в словаре, необходимо нормализовать слова во всех текстах и все переводы. Обычно приводимый в словаре перевод — это одно слово, но если приводится

словосочетание, то нормализуются все его компоненты. Если слово отсутствует в словаре, то оно оставляятся без изменений и проверятся его наличие в другом тексте в том же самом виде. Обычно такие слова — имена собственные. В некоторых случаях это помогает. Однако часто даже имена собственные переводятся, например, *Майкл-Мигель-Михаил*, или *Питер-Педро-Петр*.

Для испанского языка мы использовали систему AGME[2] (Гельбух и Сидоров, 2005), которая позволяет обрабатывать словоформы для примерно 26,000 испанских лексем.

Для английского языка была применена система, разработанная на основе тех же принципов на базе словаря WordNet, позволяющая анализировать формы для примерно 60,000 лексем.

Заметим, что морфологическая омонимия не разрешается, а в качестве возможных вариантов перевода рассматриваются имеющиеся в словаре переводы всех возможных морфологических омонимов.

Еще одна важная деталь связана с тем, что служебные слова не учитываются при подсчете, потому что они могут слишком часто повторяться, что дает высокую вероятность случайных совпадений.

Имея указанную меру сходства и функцию, которая ее вычисляет, можно построить алгоритм выравнивания пар текстов на уровне абзацев и предложений. В данный момент мы реализовали упрощенный вариант такого алгоритма, который рассматривает тройки возможных соответствий структурных единиц. Берутся три первых элемента из одного текста и три первых элемента из другого текста. Для первого элемента из первого текста вычисляется его сходство с первым, со вторым, с тертьим, с объединенным первым и вторым, с объединенным вторым и третьим и с объединенным первым, вторым, и третьим. Затем первый злемент первого текста объединяется со вторым, и процедура сравнения повторяется. Потом первый элемент объединяется со вторым и третьим, и опять повторяется вся процедура сравнения. Из полученных значений выбирается значение с максимальным соответствием.

Этот алгоритм не является оптимальным, потому что работает локально, и если случайно возникло положительное соответствие, то алгоритм ошибается в выравнивании.

Для того, чтобы дальнейшие результаты обработки не так сильно зависили от одной возможной ошибки, мы также реализовали в нашем методе известный прием контрольных точек (anchor points). Он состоит в том, что сначала определяются контрольные точки, которые обычно являются короткими кусками текста, величина сходства между которыми очень высока. Таким образом, алгоритм рботает в два прохода: сначала определяются контрольные точки, а затем производится выравнивание пар фрагментов текстов между этими точками.

В дальнейшем мы планируем использовать алгоритм, выбирающий оптимальное распределение фрагментов тектов глобально. В работе (Gelbukh *et al*., 2005) в похожей ситуации для нахождения наилучшего решения мы использовали генетический алгоритм. Другая возможность состоит в использовании техники генетического программирования.

Для оценки результатов работы алгоритма мы провели следующий эксперимент. Было произведено выравнивание для случайно выбранного фрагмента текста *Dracula*, состоящего из 50 абзацев. Среди них встречались довольно трудные случаи — напомним, что это последовательности коротких абзацев. Система выровняла правильно 94% абзацев.

Приведем пример ошибки в работе системы. Для следующей английской фразы:

«*Suppose that there should turn out to be no such person as Dr. Fergusson?» exclaimed another voice, with a malicious twang.*

*(*Букв. *«Преположим что выяснится, что нет такой особы как д-р Фергюссон?» — воскликнул другой голос, со зловещим выговором.)*

имеется такой испанский перевод

¿Y si el doctor Fergusson no existiera? -preguntó una voz maliciosa.

(Букв. *«А что если доктор Фергюссон не существует?» — спросил зловещий голос.*)

Как видим, в буквальных переводах повторяются только три знаменательных слова: *зловещий*, *голос*, *Фергюссон*. Последнее слово. будучи именем собственным, не будет найдено в словаре, но оно присутствует в том же виде в другом тексте. Интересно, что слово *спросить* в другом тексте переводится как *воскликнуть*. Это наводит нас на мысль в будущем использовать в алгоритме существующие словари, где указаны тезаурусные отношения, типа WordNet.

[2] Разработанный нами морфологический анализатор-синтезатор для испанского языка (а также аналогичную систему для русского языка, доступную для бесплатного использования в некоммерческих приложениях и включающую 100,000 лексем) можно загрузить с сайтов www.cic.ipn.mx/~sidorov и www.Gelbukh.com.

Табл. 1. Корпус англо-испанских текстов

Автор	Название
Carroll, Lewis	*Alice's adventures in wonderland*
Carroll, Lewis	*Through the looking-glass*
Conan Doyle, Arthur	*The adventures of Sherlock Holmes*
James, Henry	*The turn of the screw*
Kipling, Rudyard	*The jungle book*
Shelley, Mary	*Frankenstein*
Stoker, Bram	*Dracula*
Ubídia, Abdón	*Advances in genetics**
Verne, Jules	*Five weeks in a balloon*
Verne, Jules	*From the earth to the moon*
Verne, Jules	*Michael Strogoff*
Verne, Jules	*Twenty thousand leagues under the sea*

* Это художественный текст, а не научный.

Корпус переводных тектов

Для проверки применимости предлагаемого метода нам были нужны пары переводных текстов, в нашем случае испанских и английских. Мы построили параллельный корпус, состоящий из двенадцати текстов, см. табл. 1.

Тексты были найдены в Интернете. Все они были в формате PDF и при переводе их в текстовый формат терялась информация, а именно, конец каждой строки воспринимался как конец абзаца. Эту проблему пришлось решить вручную.

Общий объем корпуса составляет более одиннадцати мегабайт. Это может показаться очень небольшой величиной, но напомним, что речь идет о корпусе параллельных текстов, которые довольно трудно разыскать и которые требуют достаточно большой ручной работы по предватительной подготовке.

Выбор текстов Жюля Верна может показаться странным, потому что тексты на обоих языках являются переводами. В некотором смысле, это не очень важно для исследований над параллельными текстами, потому что не изучается стиль какого-либо автора. Напомним также, что получить доступ к переводным текстам весьма непросто.

Корпус доступен бесплатно для некоммерческого использования.

Выводы

В статье была рассмотрена проблема автоматической обработки пар переводных текстов, также часто называемых параллельными текстами. Одна их наиболее важных задач состоит в установлении соответствий между текстами на уровне абзацев, предложений и отдельных слов. Заметим, что при переводе это соответствие очень часто не является однозначным, т.е. одному абзацу могут соответствовать несколько, какие-то слова просто опускаются или заменяются очень дальними синонимами или фразеологическими оборотами, которые могут быть абсолютно различными в разных языках, и т.д.

Мы рассмотрели метод автоматического установления таких соответствий основанный на идее, что большинство слов все-таки должно иметь перевод, даваемый для слова в двуязычном словаре. Обычно, такие методы применяются к специализированным текстам, и неочевидно насколько они применимы для художественных текстов. Был описан один из возможных алгоритмов, реализующий данный метод. Приведены результаты экспериментов для англо-испанских текстов на уровне абзацев, которые показали точность 94%, что показывает, что методы, основанные на словарях, применимы также и к художественным текстам.

В будущем мы планируем использовать более универсальный алгоритм с глобальной оптимизацией, основанный на генетическом подходе, а также использовать в качестве дополнительного источника информации словари, в которых указываются тезаурусные отношения.

Кроме того, мы планируем работу над алгоритмом по выравниванию на уровне слов, для чего требуется информация о синтаксической структуре предложений.

Литература

1. *Гельбух, А.Ф., Г.О. Сидоров. К вопросу об автоматическом морфологическом анализе флективных языков. // Труды межд. конференции Диалог-2005, М., 2005, стр. 92-96.*
2. *Brown, P. F., Lai, J. C. & Mercer, R. L. Aligning Sentences in Parallel Corpora. // Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, 1991, pp 169 – 176.*
3. *Chen, S. Aligning sentences in bilingual corpora using lexical information. // In: Proceeding of ACL-93, 1993, pp. 9-16.*
4. *Kit, Chunyu, Jonathan J. Webster, King Kui Sin, Haihua Pan, Heng Li. Clause alignment for Hong Kong legal texts: A lexical-based approach. // International Journal of Corpus Linguistics 9:1, 2004, pp. 29–51.*
5. *Gale, W. A. & Church, K. W. A program for Aligning Sentences in Bilingual Corpora. // Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics, Berkeley, California, 1991.*
6. *Gelbukh, A. and G. Sidorov. Approach to construction of automatic morphological analysis systems for inflective languages with little effort. // Lecture Notes in Computer Science, N 2588, Springer-Verlag, 2003, pp. 215–220.*
7. *Gelbukh, A, G. Sidorov, SangYong Han. On Some Optimization Heuristics for Lesk-Like WSD Algorithms. // Lecture Notes in Computer Science, N 3513, Springer-Verlag, 2005, pp. 402–405.*
8. *Kay, Martin and Martin Roscheisen. Text-translation alignment. // Computational Linguistics, 19(1), 1993, 121-142.*
9. *Langlais, Ph., M. Simard, J. Veronis. Methods and practical issues in evaluation alignment techniques. // In: Proceeding of Coling-ACL-98, 1998.*
10. *McEnery, A. M. and Oakes, M. P. Sentence and word alignment in the CRATER project. // J. Thomas & M. Short (eds), Using Corpora for Language Research, London, 1996, pp. 211 – 231.*
11. *Meyers, Adam, Michiko Kosaka, and Ralph Grishman. A Multilingual Procedure for Dictionary-Based Sentence Alignment. // In: Proceedings of AMTA'98: Machine Translation and the Information Soup, 1998, pp. 187-198.*
12. *Mikhailov, M. Two Approaches to Automated Text Aligning of Parallel Fiction Texts. // Across Languages and Cultures, 2:1, 2001, pp. 87 – 96.*