ПОДХОД К АВТОМАТИЧЕСКОМУ ПОИСКУ ПЕРЕВОДНЫХ СЛОВОСОЧЕТАНИЙ НА ОСНОВЕ СИНТАКСИЧЕСКОЙ ИНФОРМАЦИИ И МНОГОУРОВНЕВОЙ ФИЛЬТРАЦИИ

В.И. Новицкий,
сотрудник компании ABBYY, аспирант кафедры распознавания изображений и обработки текста Московского физико-технического института (МФТИ)

Адрес: г. Москва, ул. Отрадная, д. 2Б, стр. 6
E-mail: nov.valerij@gmail.com

В работе описывается подход к автоматическому построению списка словосочетаний по корпусу выровненных параллельных текстов (текстов и их переводов, сопоставленных по цпредложениям). Подход основан на сопоставлении деревьев синтаксического разбора предложений. Ключевой особенностью данной работы является набор фильтров для удаления нерелевантных словосочетаний.

Ключевые слова: обработка естественного языка, переводные словосочетания.

1. Введение

Данная работа посвящена разработке алгоритма поиска словосочетаний вместе с их переводами на другой язык по корпусу параллельных текстов. Нас будет интересовать возможность получения нетривиальных переводов (т.н. переводных словосочетаний).

Такие словосочетания представляют собой ценный лингвистический ресурс. Например, они могут использоваться для машинного перевода текста, в качестве статистических данных в других задачах или как справочный материал для лингвистов.

Лингвистические технологии по автоматическому анализу и синтезу текстов, в свою очередь, являются полезным инструментом для автоматизации и оптимизации различных бизнес-процессов. Это может быть семантический поиск, автореферирование, машинный перевод и другие задачи, связанные с обработкой больших текстовых баз.

Особенностью данной работы является предложенный набор эвристических фильтров для выделения семантически значимых словосочетаний среди всех встретившихся в корпусе.

1.1. Постановка задачи

Дан корпус параллельных текстов (текстов и их переводов на другой язык). Тексты выровнены по предложениям. Требуется найти по ним словосочетания и их переводы.

Словосочетание — это устойчивое выражение из двух или более слов, связанных семантически (по смыслу) и грамматически.

Оба этих свойства будут использованы для поиска словосочетаний. Семантическую связь будем оценивать по частотности — слова должны достаточно часто встречаться вместе. Грамматическую связанность будем определять, анализируя дерево синтаксического разбора предложения.

Для выделения семантически-значимых словосочетаний используется ряд фильтров, описанных далее и позволяющих убрать «шумовые» высокочастотные элементы. На получаемые словосочетания дополнительно накладываются ограничения, такие, как разница в длине словосочетания и его перевода, наличие словарных соответствий между ними (или контекстами, в которых они были получены) и т.д. Данные ограничения будут подробно описаны далее.

В нашем распоряжении имеется корпус параллельных текстов. Для поиска словосочетаний используем следующую эвристику. Сопоставим деревья синтаксического разбора параллельных предложений и будем учитывать не только частоту появления той или иной комбинации слов, но то, как она переводится на другой язык. Предположение состоит в том, что словосочетание переводится на другой язык чаще всего одинаково (например, «гаечный ключ» — «wrench»).

Данная работа преследует следующие цели:

1. Разработка алгоритма поиска переводных словосочетаний (с учётом некоторых ограничений, описанных ниже).

2. Получение статистических данных (словосочетаний) для улучшения работы синтаксического анализатора (используемого в том числе в данной работе).

3. Расширение переводного словаря за счёт нахождения новых переводных статей.

4. Создание TM-базы (Translation memory) словосочетаний для использования лингвистами в качестве справочного материала.

1.2. Известные подходы к решению задачи

Существует ряд известных работ, посвящённых извлечению словосочетаний. В первую очередь следуют упомянуть работы F.A. Smadja [1, 2], считающиеся классическими в этой области. В их основе лежит статистический подход. Словосочетания порождаются для слов, часто встречающихся совместно и в определённых позициях друг относительно друга.

Высокая частота совместной встречаемости слов, как оказалось, не позволяет утверждать об устойчивости словосочетания (например, в силу специфики корпуса). Поэтому было разработано большое количество метрик, определяющих «меру ассоциации» (от англ. «association measures», или силу связанности) слов коллокации друг с другом. Описание наиболее известных мер есть, например, в [3]. В основном, эти меры учитывают частоту совместной встречаемости слов коллокации и частоту слов в отдельности по корпусу. Несколько примеров таких метрик: взаимная информация (Mutual Information, MI) [4], t-score, логарифмическая функция правдоподобия (log-likehood) [5].

Использование различных чисто статистических подходов ([1, 2, 6]) можно в первую очередь объяснить их простотой и отсутствием общедоступных и, в то же время, достаточно надёжных синтаксических анализаторов. Описываемый в данной работе подход в свою очередь базируется на использовании синтаксического анализатора, разработанного в компании ABBYY и имеющего достаточно хорошую точность. Это позволяет рассматривать предложение не как случайный поток слов, а как дерево, определяющее связи между словами предложения. В этом случае словосочетание представляет собой подграф, для которого мы знаем не только слова, входящие в словосочетание, но и зависимости между ними («меловой период» — главным является «период», подчинённым определением — «меловой»). Это позволяет нам задать лингвистические критерии фильтрации «шумовых» (неинтересных нам) словосочетаний (например, с сочинительной связью[1]). Мы можем определить понятие вложенности словосочетаний в терминах теории графов, а не теории множеств («союз республик» является вложенным словосочетанием по отношению к «союз советских республик», но не является подмножеством словосочетания «республика в союзе»).

[1] Под сочинительной связью понимается такая связь, при которой отсутствует грамматическая зависимость одного компонента синтаксической конструкции от другого компонента [7]. Например, перечисления через союзы «и», «или».

Часто для уточнения семантической значимости получаемых словосочетаний применяют наложение синтаксических шаблонов (например, «существительное+прилагательное»). Словосочетания, не подпадающие ни под один известный шаблон, считаются случайными и выбрасываются. Такой подход применяется, например, в уже упомянутых работах Frank A. Smadja [2]. В данной работе ищутся словосочетания различной длины, поэтому задача создания полной системы синтаксических шаблонов является довольно трудоёмкой. На данный момент от такого подхода решено отказаться.

Кроме того, существуют подходы, использующие синтаксическую информацию (см., например, [3]). Однако в них речь чаще всего идёт о словосочетаниях, состоящих всего из двух слов. Нас же интересуют многословные словосочетания, при поиске которых возникает проблема удаления «частичных» словосочетаний. Данная работа предлагает новый подход к её решению.

1.3. Используемые средства

В работе используются следующие алгоритмы и данные, разработанные ранее в компании ABBYY:

1. Переводной словарь (русско-английский).
2. Синтаксический анализатор (парсер).
3. Алгоритм пословного сопоставления предложений.

В основе переводного словаря лежат семантические инварианты (межъязыковые статьи). Для каждого из языков описаны различные возможные реализации этих инвариантов — синонимы (например, «бегемот» и «гиппопотам» будут лежать в одном классе). В то же самое время, омонимы будут принадлежать сразу нескольким статьям (слово «bank» будет относится и к финансовому учреждению, и к подводной мели). Задача разрешения омонимии производится на этапе анализа текста и выходит за рамки данной работы. Используемый словарь достаточно полный и состоит из более чем 60 000 семантических классов.

Синтаксический анализатор строит (как правило) несколько различных вариантов деревьев синтаксического разбора каждого предложения[2]. Для наших целей извлечения словосочетаний используется лучшее дерево синтаксического разбора (на основе внутренних оценок качества деревьев, получающихся при разрешении омонимии). В вершинах этого дерева расположены семантические инварианты, рёбра дерева — связи («главное-подчинённое»). Алгоритм может ошибаться и возвращать неправильное дерево (с неверно разрешённой омонимией). В этом случае мы полагаем, что либо словосочетания не будут порождены совсем (деревья на разных языках слишком сильно отличаются), либо будут порождены неправильные (а значит редкие) переводные словосочетания, которые будут удалены при фильтрации.

Пословное сопоставление (выравнивание) — алгоритм, ставящий в соответствие друг другу слова параллельных фрагментов текста. Данный алгоритм во время своей работы использует результаты синтаксического разбора предложения.

2. Описание применяемого подхода

Поиск словосочетаний можно разделить на следующие этапы :

1. Пословное сопоставление предложений.
2. Генерация одноязычных словосочетаний по деревьям синтаксического разбора.
3. Генерация переводных словосочетаний.
4. Фильтрация кандидатов с учётом частоты их появления в корпусе текстов.
5. Сортировка полученных результатов (словосочетания и новые переводы для словаря).

Ниже рассмотрим каждый этап подробнее.

2.1. Пословное выравнивание предложений

На данном этапе производится сопоставление слов двух параллельных предложений. Ищется наилучшее возможное паросочетание – каждому слову одного предложения ставится в соответствие не более одного слова параллельного предложения (соответствие может быть найдено не для всех слов). Для этой операции используется двуязычный словарь. Для лучшего сопоставления также учитываются зависимости в деревьях синтаксического разбора параллельных предложений.

2.2. Генерация одноязычных словосочетаний

Наложим следующие ограничения на словосочетания (и их переводы):

- ✦ Количество значимых (неграмматических) слов в пределах от одного до пяти ($l_{max} = 5$).
- ✦ Слова образуют поддерево в дереве разбора предложения.

[2] Различные представления синтаксической структуры предложения описаны, например, в [8].

✦ Среди слов словосочетания нет местоимений.

✦ Вершиной синтаксического дерева словосочетания не может быть грамматическая часть речи (стоп-слово).

✦ Разница в количестве значимых (неграмматических) слов по сравнению с переводом — не больше одного.

✦ Не более одной «дырки»[3] в словосочетании ограниченной длины.

Будем идти по дереву разбора и строить всевозможные сочетания слов, удовлетворяющие описанным выше критериям.

2.3. Генерация переводных словосочетаний

Воспользуемся результатами, полученными на двух предыдущих шагах — пословным выравниванием параллельных предложений и множеством вариантов словосочетаний, полученных по этим же предложениям. Найдём на основе этих данных всевозможные переводные словосочетания (точнее, кандидатов на роль переводных словосочетаний). Наложим на них следующие ограничения:

✦ Разница в длине словосочетания и его перевода (без учёта грамматических частей) должна быть не больше одного слова.

✦ Наличие пословных соответствий среди слов словосочетания (чем длиннее словосочетание, тем больше должно быть соответствий).

✦ Для коротких словосочетаний (1-2 слова) пословных соответствий может не быть, но тогда должны совпадать предок корня и все исходящие вершины в дереве синтаксического разбора.

✦ Если удалось сопоставить входящие или исходящие вершины для одного из словосочетаний, то эти связи должны соответствовать соответственно входящим или исходящим вершинам второго словосочетания.

На этом этапе порождаются всевозможные словосочетания. На корпусе текстов размером ≈ 4,2 млн. фрагментов (параллельных предложений) получается порядка 107 млн. различных переводных словосочетаний, из которых только около 7 млн. встречаются 2 и более раза.

2.4. Фильтрация кандидатов

На предыдущем шаге было получено большое количество словосочетаний и их переводов, встречающихся в корпусе параллельных текстов. Теперь возникает задача отобрать из них семантически значимые и устойчивые. Будем полагаться на следующий принцип: устойчивые словосочетания (почти) всегда переводятся одинаково.

Будем отбрасывать словосочетания, переводящиеся по-разному в разных документах, а также редкие (встретившиеся всего несколько раз в корпусе). Некоторые другие эвристики будут описаны ниже.

Выделим следующие этапы фильтрации полученных ранее словосочетаний:

1. Удаляем редкие словосочетания (предварительная фильтрация по частоте).
2. Удаляем словосочетания, содержащие стоп-слова.
3. Удаляем вложенные словосочетания (например, «объединённых наций»— часть словосочетания «организация объединённых наций»), если их частота не сильно превышает частоту объемлющего словосочетания.
4. Аналогично удаляем «объемлющие» словосочетания, если они встречаются значительно реже вложенных («глава организации объединённых наций»).
5. Из множества неоднозначных переводов выбираем наиболее вероятные. При этом проверяем, чтобы выбранный вариант был доминирующим (составлял не менее 70% от возможных переводов словосочетания).
6. Снова удаляем редкие словосочетания, но уже с большим порогом.
7. Удаляем известные пословные словарные переводы.
8. Сортируем полученные словосочетания.

Рассмотрим каждый из фильтров подробнее.

2.4.1. Удаление редких словосочетаний

На этапе генерации словосочетаний по деревьям синтаксического разбора мы получаем огромное количество вариантов. Большинство из них — случайные комбинации слов, не представляющие интереса. Исключим их, используя частотный фильтр. Установим порог $f_{\min}$, который определяет минимальную частоту словосочетания. Посредством анализа результатов при варьирования величины $f_{\min}$ было найдено оптимальное значение $f^*_{\min} = 5$.

2.4.2. Удаление словосочетаний, содержащих стоп-слова

Стоп-словами называются слова, запрещённые в силу каких-либо обстоятельств. Например, в дан-

[3] «Дыркой» назовём разрыв в линейном представлении словосочетания в предложении. Её размер измеряется в количестве слов.

ной работе к ним относятся числа, названия денежных единиц. Стоп-слова делятся на две категории:

1. Запрещённые в любой позиции словосочетания.
2. Запрещённые в качестве корня.

К первой категории относятся, например, числа («вторая конференция», «двухтысячный год», «два с половиной»). Ко второй — названия денежных единиц и некоторые другие классы слов.

Жёсткое удаление словосочетаний, в которых встретилось стоп-слово, может привести к потере интересных нам переводов. Поэтому фильтрация происходит только если стоп-слово встретилось в обеих частях словосочетания.

2.4.3. Удаление вложенных словосочетаний

Алгоритм генерации словосочетаний по предложению порождает все возможные их варианты. Например, в предложении «Десятое заседание Организации Объединённых наций», наряду с правильным словосочетанием «Организация Объединённых наций» мы так же получим варианты «организация наций» и «объединённых наций». Эти варианты корректны с точки зрения условий генерации словосочетаний, но не представляют интереса (в силу своей несамостоятельности).

В обратную сторону: среди переводов названия месяца «November» встречается словосочетание «ноябрь месяц», которое в общем случае не является правильным. Будем удалять и такие словосочетания.

В силу большого числа словосочетаний, попарное их сравнение для выяснения вложенности не представляется возможным. Основная идея данного этапа – введение вспомогательной структуры. Для каждого переводного словосочетания стоится множество вложенных в него «подсловосочетаний». Например, из словосочетания «первый полёт на Луну» – «first moonflight» получатся «полёт на луну» – «moonflight», «первый полёт» – «first flight» и т.п. Каждому получившемуся вложенному словосочетанию приписывается частота его «родительского» словосочетания. Затем одинаковые «вложенные» словосочетания (от разных «родителей») объединяются в одно, а частотой получившегося словосочетания берётся максимальная из частот. Как результат, мы получаем новый список из словосочетаний, который можно легко сравнить с исходным. Для удобства назовём этот список массивом «кусочков».

Если для словосочетания удаётся найти пару (такое же) в массиве «кусочков», это означает, что есть (минимум одно) словосочетание большего размера, для которого данное является вложенным. Нам известна максимальная частота «большего» словосочетания (как атрибут «кусочка»). Если она достаточно велика (не сильно меньше частоты рассматриваемого словосочетания), то рассматриваемое словосочетание является неотъемлемой частью некоего большего по размеру и должно быть удалено. Таким образом, удаётся избавиться от «вложенных» словосочетаний.

Для решения обратной задачи (удаления «объемлющих» словосочетаний) необходимо модифицировать массив «кусочков», дописав в атрибуты так же ссылки на исходные словосочетания, из которых были получены кусочки, и их частоты. В этом случае можно так же оценить частоту «большего» словосочетания (соответствующего «кусочку»). Если она существенно ниже частоты рассматриваемого словосочетания, то оставить нужно именно «вложенное» словосочетание. А «внешнее» – удалить.

2.4.4. Разрешаем неоднозначности переводов

В нашей модели словосочетания — устойчивые выражения. Переводные словосочетания также обладают устойчивым переводным соответствием. Значит, для «правильного» словосочетания среди всевозможных вариантов его перевода должен быть один доминирующий.

Например, из-за неточностей перевода и ошибок разбора наряду с правильным переводом «левый» – «left» среди словосочетаний появляются соответствия «левый» – «right». Для их отсечения в основном и предназначен рассматриваемый фильтр. На этом этапе накладываются достаточно жёсткие требования – доминирующий перевод должен составлять не менее 70% всех переводов.

2.4.5. Повторная фильтрация по частоте

Для того, чтобы позволить последующим фильтрам отработать правильно на словосочетаниях с близкой к пороговой частотой появлений, мы временно оставляем не очень редкие, но всё же «непроходные» по нашим порогам словосочетания. На данном этапе мы их окончательно удаляем.

2.4.6. Удаление пословных переводов

Для переводных словосочетаний, каждая из частей которых состоит из одного слова, проверим, не являются ли они переводами по словарю («левый» – «left»). Удаляем их в подобных случаях.

2.4.7. Сортировка полученных результатов

Полученные в результате фильтрации словосочетания можно условно разделить на две группы:

1. Несловарные переводы слов (один к одному), которых нет в нашем словаре.

2. Собственно, переводные словосочетания.

Первая группа состоит из переводных словосочетаний с одним значимым словом в каждой из частей (условия генерации словосочетаний их допускают). Полученные таким образом соответствия служат лингвистам справочным материалом для обнаружения недостающих связей в словаре, а так же для исправления описаний.

Вторая группа содержит искомые переводные словосочетания.

2.5. Порядок применения фильтров

Рассмотренные выше фильтры выполняются в том же порядке, в котором они описаны в данной работе. Рассмотрим, какими факторами это обусловлено.

Частотный фильтр удаляет словосочетания, которые априори не представляют интереса для дальнейшего рассмотрения. Мы считаем, что если словосочетание редкое, то дальнейшими статистическими проверками мы не сможем доказать его значимость. Фильтр стоп-слов так же удаляет априори неинтересные варианты, уменьшая необходимую работу на последующих этапах. Важным также является тот факт, что после применения этих фильтров количество словосочетаний уменьшается на 90-95%.

Фильтры «вложенных»/«внешних» словосочетаний, а также неоднозначных переводов применяются на одном множестве словосочетаний, так как каждый из них требует максимально полной информации о «близких» к ним словосочетаниях (отличающихся дополнительными словами или другими переводами).

Повторная фильтрация по частоте призвана удалить те редкие словосочетания, которые мы намеренно оставили для корректной работы других фильтров (например, фильтра неоднозначных переводов). Наряду с фильтром пословных переводов они удаляют то, что не должно войти в результат по итогам фильтрации. Их порядок значения не имеет (суммарный результат остаётся одинаковым).

На этапе сортировки получаются окончательные результаты. Этот фильтр работает теми словосочетаниями, которые представляют практический интерес. Соответственно, он выполняется последним.

3. Количественные результаты

Эксперименты проводились на русско-английском корпусе, содержащем 4,2 млн. фрагментов (параллельных предложений).

Алгоритм порождения переводных словосочетаний даёт 62 млн. различных пар (всего порождается около 107 млн. словосочетаний). Из них подавляющее большинство (56 млн.) встречаются только один раз. Динамика количества остающихся словосочетаний на различных этапах фильтрации показана *в табл.* 1.

Таблица. 1.

Динамика фильтрации словосочетаний

Название фильтра	Словосочетаний на выходе
По частоте (предварительно)	2,5 млн.
По списку стоп-слов	1,1 млн.
Вложенные словосочетания	568 тыс.
Неоднозначные переводы	105 тыс.
Словарные переводы	66,5 тыс.
По частоте (окончательно)	42 636
Сортировка результатов	42 535

Таким образом, на экспериментальном корпусе удаётся получить ≈ 42,5 тыс. переводных словосочетаний. Несколько примеров новых переводов (отсутствовали в используемом словаре), найденных на этом этапе приведены *в табл. 2.*

Таблица. 2.

Примеры новых словарных связей (отсутствовали в словаре)

Английский	Русский
Glib	Бойкий
Ledger	Гроссбух
Reciter	Чтец

Несколько примеров из найденных словосочетаний приведены *в табл. 3.*

Таблица. 3.

Примеры найденных словосочетаний

Английский	Русский	Встретилось в корпусе
job time	срок задания	12
galaxy space	космическое пространство	13
other foreign object	иной посторонний объект	5
air transport field	область воздушного транспорта	30
to be beyond the scope of book	выходить за рамки книги	29
to establish under article	учреждать в соответствии со статьёй	75

Существуют две общепринятые меры измерения качества полученных результатов — точность и полнота. В данной работе основной упор на повышении качества получаемого результата. Это связано с тем, что практически невозможно вручную просмотреть несколько десятков тысяч словосочетаний для поиска в них ошибок. Полнота же может быть увеличена, например, за счёт увеличения текстовой базы. В отличие от статистических подходов, мы предпочитаем удалять редкие словосочетания, так как достаточно высока вероятность встретить среди них те, которые получены в результате ошибок разбора или пословного сопоставления.

Для оценки качества мы используем два подхода. Первый является косвенным — проводится синтаксический разбор тестовой выборки с и без полученных словосочетаний. Сравнение полученных результатов часто позволяет найти ошибки алгоритма фильтрации (те словосочетания, которые не должны были пройти этот этап).

Второй подход — это ручная проверка лингвистами случайной выборки. Пример такой оценки представлен *в табл. 4.* Словосочетания, забракованные экспертом, условно разделены на 3 категории: ошибки словаря (лингвистического описания), ошибки алгоритма (для которых известны пути исправления) и прочие ошибки. В частности, одна из основных проблем — отсутствие учёта дублирующихся фрагментов (одинаковые предложения из инструкций, приказов и подобных шаблонных документов, имеющихся в базе). С учётом добавления этого функционала можно достигнуть качества в 80% и более.

Таблица. 4.

Экспертная оценка качества на выборке из 100 словосочетаний.

Качество	Процент
Хорошие словосочетания	67
Ошибки в описаниях	4
Недоработки алгоритма	16
Другое	12

4. Заключение

Результатом проведённого исследования является алгоритм, позволяющий находить переводные словосочетания по корпусу параллельных текстов.

Есть несколько направлений возможного дальнейшего его развития.

- ✦ Введение функции оценки качества полученных словосочетаний.
- ✦ Настройка пороговых значений для фильтров.
- ✦ Добавление новых фильтров.
- ✦ Учёт дублирования отдельных фрагментов. ■

Литература

1. Smadja F.A. Retrieving collocations from text: Xtract // Computational Linguistics. 1993 Vol. 19. No. 1. – Pp. 143-177.
2. Smadja F.A., McKeown K. Translating collocations for use in bilingual lexicons. – HLT, 1994.
3. Evert S. The Statistics of Word Cooccurences Word Pairs and Collocations. PhD thesis / Universität Stuttgart. – Institut für Maschinelle Sprachverarbeitung (IMS), 2004.
4. Church K.W., Hanks P. Word association norms, mutual information, and lexicography // Computational Linguistics. 1990. Vol. 16. No. 1. – Pp. 22-29.
5. Dunning T. Accurate methods for the statistics of surprise and coincidence // Computational Linguistics. 1993. Vol. 19. No. 1. – Pp. 61-74.
6. Bouma G. Collocation extraction beyond the independence assumption // Proceedings of the ACL 2010 Conference Short Papers / ACLShort '10. – Stroudsburg, PA, USA: Association for Computational Linguistics, 2010. – Pp. 109-114.
7. Казарина В.И. Современный русский синтаксис: структурная организации простого предложения. – Елец: ЕГУ, 2007.
8. Bolshakov I.A., Gelbukh A.F. Computational Linguistics: Models, Resources, Applications. – IPN - UNAM - Fondo de Cultura Economica, 2004.