Н. Ю. Дюмин, Т. Ю. Дюмина
N. Yu. Dyumin, T. Yu. Dyumina

МЕТОДЫ ВЫЧИСЛЕНИЯ КООРДИНАТ ВЕКТОРОВ ПРИ АВТОМАТИЧЕСКОМ ФОРМИРОВАНИИ ДВУЯЗЫЧНЫХ ЛЕКСИКОНОВ

METHODS FOR COORDINATES CALCULATION IN AUTOMATIC BILINGUAL LEXICON ACQUISITION

Аннотация. В данной статье предлагается алгоритм автоматического формирования двуязычного лексикона для систем АОТ на материале параллельного корпуса текстов патентов в рамках векторной модели. Также приводится сравнительный анализ эффективности методов построения векторов. Кратко приводятся основные проблемы алгоритма и возможные пути их решения.

Abstract. The paper presents an algorithm of automatic bilingual lexicon acquisition for NLP systems using parallel corpora within Vector Space Model. In this paper, we analyze the effectiveness of different methods for coordinate calculation. In brief, the algorithm major problems and their solutions are discussed.

Не вызывающая сомнений необходимость в автоматической обработке текстов на естественном языке особенно актуальна для технических текстов. Технические тексты, включающие в себя инструкции, патенты и т.д. представляют собой описания объектов действительности с целью сообщения информации о самих объектах, особенностях их организации, способах их использования. Это определяет необходимость использования точного и подробного описания референтов, что, в свою очередь, обуславливает первостепенную важность ИГ в этих текстах. Важно отметить, что технические тексты, например, патенты зачастую описывают новые референты, а значит, велика вероятность отсутствия тех или иных компонентов ИГ в лексиконе системы АОТ, которая реферирует, переводит или выполняет информационный поиск этих текстов. Особенно эта

проблема релевантна для систем, обрабатывающих текст по правилам.

Процесс пополнения лексикона систем АОТ, особенно многоязычных, известен своей затратностью в отношении временных и интеллектуальных ресурсов, что делает актуальной задачу автоматического формирования лексиконов.

Способы автоматического формирования лексикона

К настоящему времени существует ряд алгоритмов, позволяющих автоматически формировать как одно- так и многоязычные лексиконы. В частности, одноязычные лексиконы, описывающие лексические отношения: гипонимы\гиперонимы[1],[2] или меронимов. Приведенные системы используют шаблоны вида «X, часть Y» или «Y, в частности X». Также лексиконы словоформ, содержащие грамматические характеристики, например, морфологический лексикон по тексту на словацком языке[3]. Данная система опирается на статистические методики – для словоформ в корпусе просчитывается вероятность того, что они являются словоформами одной леммы. Статистические методики также используются для извлечения эквивалентов при составлении двуязычных лексиконов[4].

---

[1] *Hearst M.* Automatic Acquisition of Hyponyms from Large Text Corpora. Proc. of COLING 92, Nantes, 2, 1992.

[2] *Shaikevich A.* Automatic Construction of a Thesaurus from Explanatory Dictionaries, Automatic Documenation and Mathematical Linguistics, 19(2), 1985.

[3] *Sagot B.* Automatic Acquisition of a Slovak Lexicon from a Raw Corpus, TSD05

[4] *Sahlgren M., Karlgren J.* Automatic bilingual lexicon acquisition using random indexing of parallel corpora; *Tornfeldt T.* Graph similarity, parallel texts and automatic bilingual lexicon acquisition, MA Thesis.

Векторная модель при автоматическом формировании лексикона

Предлагаемый алгоритм использует векторную модель для автоматического установления эквивалентов в параллельных корпусах текстов.

При этом, весь корпус представляет собой пространство, измерениями которого являются входящие в его состав сегменты, например, отдельные тексты. Сегменты также могут быть представлены и отдельными предложениями в случае, если корпусы выровнены по предложениям.

Каждая словоформа корпуса может быть представлена в виде вектора, описывающего её дистрибуцию в корпусе. Для вычисления значения координат вектора используются различные методы, например, простейший из них, установление факта вхождения словоформы в сегмент строит бинарный вектор. Единицы такого вектора указывают на сегменты, в которых словоформа встречается один и более раз, нули указывают, что словоформа не встречается в данных сегментах.

В данной работе мы рассматриваем три способа определения координат вектора: мера TF-IDF, мера TF и частота словоформы в сегменте.

Мера TF для словоформы w вычисляется как удельный вес словоформы в сегменте $S = [x_1, x_2..x_n]$:

$$TF = \frac{count(w)}{n},$$

где count(w) – частота словоформы в сегменте, а n – общее количество словоформ сегмента.

Мера TF-IDF вычисляется как произведение мер TF и IDF, где IDF вычисляется следующим образом:

$$IDF = log \frac{|D|}{count(w, D)},$$

где |D| – общее количество сегментов, а count(w,D) – количество сегментов, в которых встречается словоформа w.

Под мерой частоты словоформы в сегменте мы понимаем количество вхождений словоформы w в сегмент S.

После того, как для каждой словоформы корпуса вычислен вектор, производится поиск ближайших векторов в параллельном корпусе, теоретическим основанием которого является предположение, что дистрибуции эквивалентов должны быть схожи. Ввиду грамматической разницы естественных языков более закономерным является приоритет ИГ.

Поиск ближайших векторов может проводиться различными способами, среди которых различные метрики, коэффициенты корреляции и угловые меры. В данном исследовании используется мера косинуса, в ходе экспериментов показавшая наилучшие результаты.

Эксперимент по выявлению эффективности различных методов вычисления координат

Для проведения эксперимента использовались параллельные корпусы технических текстов, представленных аннотациями к патентам на русском и английском языках. Тексты взяты из базы патентов Всемирной организации интеллектуальной собственности (**www.wipo.org**). Общий объём корпуса составил 168 тыс. словоупотреблений, в том числе, 87,5 тыс. словоупотреблений в английском корпусе и 80,5 тыс. в русском. Количество словоформ русского языка составило 11446 единиц, английского – 4508 единиц. В качестве сегментов выступили полные тексты аннотаций, количество – 868 сегментов. Предметная область – B61 «Железная дорога» и B62 «Наземные транспортные средства (кроме железной дороги)» по международной патентной классификации IPC.

Естественно, такая типологическая разница оказала значительное влияние на результаты эксперимента и значительно снизила эффективность формирования лексикона. Однако целью данного эксперимента в большей степени являлся сравнительный

анализ различных методик построения векторов, поэтому мы намеренно не предпринимаем попыток решения известных проблем: 1) несоответствия морфологических парадигм имени и глагола в русском и английском языках, 2) несоответствия количества компонентов многокомпонентных ИГ. Решением первой проблемы представляется лемматизация. Вторую проблему можно решить синтаксическим анализом, например, извлекая и выравнивая не словоформы, а именные группы.

Таким образом, объектами поиска эквивалентов являлись словоформы.

Эксперимент проводился в одну сторону, т.е. эквиваленты подбирались для 4508 английских словоформ.

Таблица 1. Результаты эксперимента

Мера	**Точность**	**Полнота**	**Удельный вес**
TF-IDF	932	4508	0,206
TF	934	4508	0,207
Частота	953	4508	0,211

Первоначальные результаты приведены: табл. 1. Для оценки эффективности использовались следующие показатели: точность, которую мы определили, как количество правильно установленных эквивалентов, полнота – общее количество извлеченных эквивалентов и удельный вес, в качестве интегрального коэффициента, – отношение точности к полноте.

Несмотря на ожидания, меры TF-IDF и TF не показали лучших результатов по отношению к подсчету частоты. Напротив, эксперимент показал, что корпус такого рода нивелирует значение меры IDF (результат меры TF выше, чем у меры TF-IDF: табл. 2), т.к., вероятно, определенные лексемы встречаются во многих документах, при этом являясь ключевыми или частью ключевых словосочетаний. Заметим, что имеется лишь 2 случая, когда мера TF-IDF нашла эквивалент словоформе, с которой «не справилась» мера TF.

Таблица 2. TF vs TF-IDF

Мера	**Результат**	**Близость векторов**
Tfidf	*compliance:целевого*	0.38883
Tf	compliance:подавлении	0.38883
Freq	compliance:подавлении	0.26145

Частота, в отличие от «степени важности» определяемой TF-IDF и TF показала более стабильный результат.

Отдельно следует отметить, что имеется значительное количество случаев, в которых меры TF-IDF и TF устанавливали правильный эквивалент, в то время как мера частоты давала ошибочный результат (например: табл. 3).

Таблица 3. TF-IDF, TF vs Частота

Tfidf	force:силу	0.49816
Tf	force:силу	0.49816
Freq	*force:получаемой*	0.52000

Вероятно, совместное использование двух мер (TF и меры частоты) могло бы повысить общую эффективность алгоритма. Тем не менее, к настоящему моменту установить какую-либо корреляцию не удалось, а, следует, не ясно основание для решения о том, который из вариантов является правильным эквивалентом.

Кроме того, встречаются случаи нахождения разными мерами разных словоформ одной лексемы: табл. 4. Такие случаи могут позволить собирать морфологическую информацию о вхождениях лексикона.

Таблица 4. Морфологические варианты лексемы

Tfidf	turbojet:турбореактивн*ые*	0.30965
Tf	turbojet:турбореактивн*ые*	0.30965

Freq	turbojet:турбореактивн*ых*	0.15485

Заключение

Векторная модель является перспективным способом автоматического формирования двуязычных лексиконов, т.к. представляет возможность сократить временные и интеллектуальные усилия за счёт вычислительных. Низкие результаты описанного эксперимента (точность ~20%) обусловлены специальными условиями его проведения. В частности, при введении порогового значения 0,7 для коэффициента близости векторов эффективность увеличивается до ~30%. Предварительное удаление стоп-слов также улучшает общий результат.

Помимо указанных выше решений проблем, перспектива развития алгоритма видится в привлечении новых методов вычисления координат и комбинировании результатов.