Статистическая модель для распознавания смыслов в текстах иностранного языка с обучением на примерах из параллельных текстов

© А.Е. Ермаков © П.Ю. Поляков

ООО «ЭР СИ О»,
Москва, Россия

ermakov@rco.ru pavel@rco.ru

Аннотация. Распознавание смыслов (упоминаний целевых ситуаций, событий и фактов) в текстах иностранного языка в идеале требует разработки синтаксического анализатора этого языка и ряда сопутствующих лингвистических компонентов. В работе предложен альтернативный подход к построению распознавателя смыслов, не требующий глубокого машинного анализа языка текста. Подход строит статистическую модель распознавателя смысла в форме *n*-ок совместно встречающихся слов, с возможностью вставки не более заданного количества посторонних слов между словами *n*-ок. Для обучения модели использованы корпус параллельных текстов и русскоязычный лингвистический анализатор, который выделяет целевые смыслы из русских текстов, отбирая фрагменты, релевантные смыслам, в параллельных текстах иностранного языка. Описаны результаты экспериментов по распознаванию смыслов на корпусе квазипараллельных русско-армянских новостных текстов, в том числе процедура предварительного выравнивания текстов по параллельным фрагментам.

Ключевые слова: машинный анализ текстов на иностранных языках, кросс-языковой информационный поиск, распознавание смысла в тексте, извлечение событий и фактов, статистическое машинное обучение на параллельных текстах, выравнивание параллельных текстов.

Statistical Model for Recognition of Senses in Foreign Language Texts Trained by Examples from Parallel Texts

© Alexander Ermakov © Pavel Polyakov

RCO Llc,
Moscow, Russia

ermakov@rco.ru pavel@rco.ru

Abstract. Recognition of senses (mentioning of target situations, events and facts) in foreign language texts needs developing of a syntactic analyzer and some linguistic components for this language. The alternative approach to construct a senses recognizer that does not need complex machine analysis of the language of a text is proposed in the report. This approach builds a statistical model of a senses recognizer in a form of n-tuples of words that stand together in the text, permitting insertion of a few other words between them. To train the model, a corpus of parallel texts and a Russian linguistic analyzer are applied. The linguistic analyzer is used to extract target senses from Russian texts, selecting the fragments that are relevant to these senses in parallel texts in a foreign language. The results of experiments in senses recognition in the corpus of quasi-parallel Russian-Armenian news texts are described, as well as a preliminary procedure of parallel text fragments alignment.

Keywords: machine analysis of foreign language texts, cross-language information retrieval, recognition of sense in text, events and facts extraction, statistical machine training using parallel texts, parallel texts alignment.

1 Введение

Вопросы межязыкового информационного поиска стали предметом систематического исследования уже с 1990-х годов [2]. Основные результаты и направления современных исследований отражены в работах [3–5, 7]. В центре их внимания оказались статистический машинный перевод, автоматическое построение словарей перевода слов, терминов и именованных сущностей, перевод и расширение поисковых запросов, а также формирование и выравнивание корпусов параллельных текстов-переводов как источников,

Труды XIX Международной конференции «Аналитика и управление данными в областях с интенсивным использованием данных» (DAMDID/ RCDL'2017), Москва, Россия, 10–13 октября 2017 года

необходимых для обучения всех статистических алгоритмов.

В основе предлагаемого нами подхода лежат идеи, имеющие аналогии с таковыми, используемыми в статистическом машинном переводе, наиболее полная информация по которому представлена на веб-ресурсе [7]. Тем не менее, предложенная модель и исследования, посвященные ей, нам не встречались.

Под присутствием заданного смысла в тексте будем понимать описание или упоминание в этом тексте:

- фактов и ситуаций определенного класса, например: *владение акциями предприятий*, *заключение договоров между организациями*, *встречи персон*;
- определенных событий, например: *война в Сирии*, *санкции против России*;
- определенных тем, например: *образ России в зарубежных СМИ*, *политика Дональда Трампа*.

Тогда задачу информационного поиска в общем виде можно представить как задачу распознавания присутствия заданного смысла в анализируемых текстах и выделения фрагментов текста, релевантных искомому смыслу.

Для распознавания смыслов в русскоязычном тексте можно использовать разработанный нами лингвистический анализатор RCO Fact Extractor [8], который извлекает структурированные описания ситуаций, событий и фактов, выраженные в тексте заданными конфигурациями синтаксически связанных слов [9].

Адаптация русскоязычного лингвистического анализатора к новому языку представляет собой нетривиальную ресурсоемкую задачу, требующую построения синтаксического анализатора этого языка и ряда сопутствующих лингвистических компонентов. В настоящей работе предложен альтернативный подход к построению распознавателя смыслов на иностранном языке, не требующий глубокого машинного анализа этого языка. Подход строит модель статистического распознавателя смысла на новом языке в форме *n*-ок совместно встречающихся слов, с возможностью вставки не более заданного количества посторонних слов между словами *n*-ок. Появление всех слов какой-либо из *n*-ок в пределах текстового окна ограниченной длины интерпретируется как наличие целевого смысла. На практике поиск смысла, описанного в такой форме, может быть эффективно реализован средствами поисковой машины, поддерживающей поиск заданных слов в пределах окна заданной длины с сохранением заданного порядка слов или без такового.

Для обучения распознавателя использованы корпус параллельных текстов и русскоязычный лингвистический анализатор, который выделяет целевые смыслы и содержащие их фрагменты из русских текстов на основе синтактико-семантических шаблонов [9]. Параллельные им фрагменты из текстов иностранного языка также считаются релевантными смыслам и используются для последующей настройки параметров статистической модели. Такой подход требует для настройки распознавателя на каждый новый язык: а) соответствующего параллельного корпуса, представительного в плане присутствия разных способов выражения целевых смыслов; б) простейшего лингвоанализатора, способного строить варианты нормальных форм для словоформ иностранного языка; в) для некоторых видов распознаваемых смыслов от лингвоанализатора может потребоваться умение выделять именованные сущности.

2 Модель статистического распознавателя смыслов

Будем называть *смысло-текстом* текстовый фрагмент, содержащий такую конфигурацию синтаксически связанных слов, появление которой в произвольном тексте говорит о присутствии в нем заданного смысла. Идеальным смысло-текстом является такой фрагмент, в котором отсутствуют лишние слова, появление которых не является обязательным для идентификации присутствия смысла, например: *Берлага заключил договор с Корейко; договор Берлаги и Корейко* (для смысла «*договора между персонами*»); *усиление влияния России на Ближнем Востоке*; *Газпром использует свое монопольное положение на рынке энергоносителей* (смысл «*образ России в зарубежных СМИ*»).

Определим статистический распознаватель смыслов (СРС) как механизм, который для данного текста d определяет, присутствует ли в нем заданный смысл *Se*: формирует реакцию Re(*Se*,*d*)=1, если смысл присутствует, и Re(*Se*,*d*)=0, если отсутствует.

Построим модель СРС в следующем виде. Распознаватель считает, что смысл *Se* присутствует в тексте *d* (реакция Re(*Se*,*d*)=1), если текст содержит хотя бы одну *n*-ку из множества $S=U_{g,n}\ s^g{}_n$, g=0..G, n=1..N, где $s^g{}_n=\{(w_1,w_2,\ldots,w_n,g)\}$ – подмножество *n*-ок, каждая из которых содержит n определенных слов w_i, допуская между ними вставку произвольных слов в количестве, не превышающем *g*. Далее будем обозначать профиль СРС как $S=\{s_1,s_2,\ldots,s_J\}$, где J – количество *n*-ок в профиле, нумеруя подряд *n*-ки в профиле и опуская обозначения *n* и *g* в них. Множество *n*-ок *S* будем называть профилем смысла *Se*. В зависимости от степени свободы порядка слов в языке к словам *n*-ок либо следует применять требование сохранения их порядка в окне (пр., армянский, казахский), либо нет (сербский, белорусский). С практической точки зрения достаточными представляются значения N=4, что соответствует, например, упоминанию целевого объекта с тремя дополнительными словами, достаточно точно идентифицирующими искомую ситуацию с объектом.

Обучение СРС смыслу *Se* представляет собой процедуру поиска такого профиля *S*, который обеспечит наилучшее качество работы СРС на

текстах обучающего корпуса D.

За оценку правдоподобия профиля S возьмем совокупную оценку ожидаемых от него полноты P и точности R распознавания смысла (т. н. F_1-мера в теории информационного поиска):

$$q(S,D)=2\mathrm{P}(S,D)\mathrm{R}(S,D)/(\mathrm{P}(S,D)+\mathrm{R}(S,D)), \quad (1)$$

где $\mathrm{P}(S,D)=|D^*_1(S)|/|D_1(\mathrm{S})|$, $\mathrm{R}(S,D)=|\mathrm{D}^*_1(\mathrm{S})|/|D(Se)|$, $D(Se)$ – множество смысло-текстов обучающего корпуса, релевантных смыслу Se, $D_1(S)$ – множество всех смысло-текстов, распознанных профилем S, $D^*_1(S)$ – множество смысло-текстов, правильно распознанных профилем S. Тогда наилучший профиль S^*, обеспечивающий максимальное качество СРС, определится как:

$$S^*=\arg\max_S q(S,D) \quad (2)$$

Для ускорения поиска максимума $q(S,D)$ в пространстве S комбинаций n-ок определим правдоподобие вхождения отдельной n-ки s_j в S^* как

$$q(s_j)=(1-1/|D^*_1(s_j)|D^*_1(s_j)|/|D_1(s_j)|, \quad (3)$$

где множитель $|D^*_1(s_j)|/|D_1(s_j)|$ характеризует ожидаемую точность, а множитель $1-1/|D^*_1(s_j)|$ повышает вероятность включения в профиль n-ок с большей частотой встречаемости в релевантных смысло-текстах $D^*_1(s_j)$, поскольку от таковых ожидается большая полнота распознавания смысла. Тогда наилучший профиль S^* в соответствии с (2) можно построить, применив следующий жадный алгоритм поиска в пространстве состояний.

Вначале алгоритм собирает все уникальные n-ки s_j, для которых значение $q(s_j)$ в соответствии с (3) выше определенного порогового значения – кандидатов на включение в профиль. Каждой n-ке-кандидату соответствует массив идентификаторов содержащих ее смысло-текстов d_i.

Далее n-ки сортируются по убыванию значений $q(s_j)$, и первая n-ка включается в профиль на шаге 1: $S_1=\{s_1\}$, чем начинается выполнение итерационного алгоритма расширения профиля новыми n-ми, идя по убыванию значений $q(s_j)$. Обозначим S_{t-1} профиль, полученный на итерации t–1, а s_{t-1} – последнюю обработанную n-ку, включенную или не включенную в профиль. На следующей итерации t производится попытка добавить к профилю очередную n-ку s_t. Вычисляются оценки качества нового получаемого профиля $\mathrm{P}(S_t,D)$, $\mathrm{R}(S_t,D)$, $q(S_t,D)$, и новый профиль S_t признается лучше старого при одновременном соблюдении следующих условий:

$$q(S_t,D)>q(S_{t-1},D) \text{ и } \mathrm{RG}(s_t|S_{t-1},D)/\mathrm{TG}(s_t|S_{t-1},D)>\mathrm{P}_{\min}, \quad (4)$$

где $\mathrm{P}_{\min}$ – минимальная допустимая точность профиля (мы использовали $\mathrm{P}_{\min}=0.7$), $\mathrm{RG}(s_t|S_{t-1},D)$ – прирост количества релевантных смысло-текстов, распознаваемых профилем S_{t-1} после добавления к нему n-ки s_t, $\mathrm{TG}(s_t|S_{t-1},D)$ – прирост количества всех смысло-текстов, распознаваемых профилем S_{t-1} после добавления к нему n-ки s_t.

При выполнении обоих условий n-ка добавляется к профилю S_{t-1}, и формируется новый профиль S_t, который принимается за S^*; в противном случае n-ка пропускается, и делается попытка добавления к профилю следующей n-ки s_{t+1} – итерация t+1. Расширение профиля прекращается при прохождении всех n-ок-кандидатов или при достижении порога по допустимому количеству n-ок в профиле. Тогда производится возвращение на шаг назад к профилю без добавления последней n-ки, делается попытка добавить следующую за ней n-ку из числа кандидатов и т. д. Таким способом обходится дерево возможных комбинаций n-ок в профиле, и наилучший полученный профиль S^* запоминается. При включении n-ок в порядке убывания их $q(s_j)$ можно ожидать, что лучшие варианты профиля будут получены на более ранних шагах алгоритма.

3 Выравнивание параллельных текстов

Для обучения СРС необходимо сформировать обучающее множество смысло-текстов $D(Se)=\{d_i\}$, релевантных смыслу Se. В качестве таковых отбираются смысло-тексты иностранного языка, параллельные тем русскоязычным смысло-текстам, в которых лингвистическим анализатором выделен смысл Se. Источниками параллельных смысло-текстов, достаточно объемными и представительными в плане разнообразия содержания, являются корпуса переводов новостных сообщений. Такие корпуса в общем случае не содержат строго параллельных текстов, в которых предложения с одинаковыми порядковыми номерами в последовательности могли бы выступать в роли параллельных смысло-текстов. Более того, переводы новостных сообщений часто содержат иную разбивку на предложения, чем их оригиналы, в том числе нередко опускают оригинальные предложения и вставляют новые. Аналогично, перевод предложения может содержать пропуски/вставки ряда значимых слов в описании ситуации – переводчики новостей нередко опускают детали или добавляют собственные интерпретации.

Вследствие этого обучение профилей СРС требует проведения машинной процедуры предварительного выравнивания квазипараллельных текстов, которая устанавливает соответствие между предложениями на двух языках по принципу «одно к одному», «одно к нескольким» или «несколько к одному», а также отбрасывает предложения, перевод которых является излишне «вольным».

Обычно методы выравнивания предложений используют алгоритм динамического программирования, который позволяет вычислительно эффективно определить такую последовательность пар сопоставленных друг другу предложений, для которой сумма расстояний между предложениями в каждой паре будет минимальна. При этом сущность используемого метода заключается в способе определения сходства между парой предложений двух языков. В качестве русскоязычной точки входа в методы выравнивания можно указать работу отечественных исследователей [11]. Наиболее полная информация с зарубежной библиографией по данной теме доступна на веб-ресурсе [6].

Реализованный нами метод требует наличия словаря переводных соответствий слов двух языков, желательно с вариантами синонимичных переводов, а также лингвистических анализаторов обоих языков, способных разделять текст на предложения, а предложения – на сущности, которым приписываются варианты их перевода. В качестве сущностей анализаторы должны выделять слова и, желательно, словосочетания, обозначающие различные классы именованных (персоны, организации, географические объекты) и специальных (даты, периоды времени, денежные суммы) объектов. Именованные и специальные сущности в новостных текстах являются опорными точками для выравнивания параллельных предложений. Сущности приписывается набор альтернативных вариантов перевода (если это удается), а в некоторых случаях, например, для именованных персон, – еще и вариант транскрибирования.

Будем называть количеством сопоставлений переводов $Eq(e_i|d_j)$ количество сущностей из предложения $e_i=\{e^k_i\}$, k=1..K, i=1..I, сопоставленных с сущностями из предложения $d_j=\{d^p_j\}$, p=1..P, j=1..J. Здесь I и J – количества предложений в параллельных текстах E и D; K и P – количества сопоставляемых сущностей в соответствующих предложениях i и j, из числа которых исключены общеупотребимые слова обоих языков, вероятность совпадения переводов которых в паре произвольных предложений высока (прежде всего, это союзы, местоимения, предлоги).

Сущности e^k_i и d^p_j считаются сопоставленными, если выполняется любое из трех условий:

1. обе сущности относятся к классу специальных, и их тип (дата, период времени, денежная суммы) одинаков;
2. один из вариантов имени сущности точно совпадает с одним из вариантов перевода/транскрипции одного из имен другой сущности;
3. условие 2 выполняется не для точного, а для «нечеткого» совпадения, когда эквивалентными признаются строки, имеющие относительное количество совпавших триграмм символов не менее порогового.

Условие 1 позволяет сопоставить сущности, выражаемые специальными конструкциями (пр. даты), для которых получение совпадающих переводов маловероятно вследствие разнообразия используемых форматов написания в каждом из языков.

Условие 3 необходимо для сопоставления, в первую очередь, именованных сущностей – персон и организаций, при переводе которых человеком-переводчиком часто не соблюдается исходный формат, кроме того, в силу потенциальной неполноты словарей перевода имен, не все части сложных имен могут иметь варианты перевода в словаре. Так, имена персон (как полные, так и краткие) обычно удается сопоставить именно по «нечеткому» совпадению транскрипций. Нередко такое сравнение транскрипций работает для географических мест, обычно не общеизвестных (местных), а также для организаций, напротив, общеизвестных (международных).

Заметим, что величина $Eq(e_i|d_j)$ и вычисляемая наоборот величина $Eq(d_j|e_i)$ в общем случае будут иметь различные значения в силу возможных повторений слов или вариантов их переводов в одном предложении, а также в силу использования «нечеткого» сравнения строк.

Мера прямого сходства переводов определяется как $Tr(e_i|d_j)= Eq(e_i|d_j)/K$, а мера обратного сходства – как $Tr(d_j|e_i)=Eq(d_j|e_i)/P$.

Обозначим $(i(t), j(t))$, t=1...T, последовательность номеров пар предложений e_i и d_j из параллельных текстов $E=\{e_i\}$, i=1..I, и $D=\{d_j\}$, j=1..J, где $j(1)\geq 1$, $i(1)\geq 1$, $j(T)\leq J$, $i(T)\geq I$. Здесь t – переменная, введенная для установления возможного соответствия между номерами предложений $i(t)$ и $j(t)$. Тогда $(i(t),j(t))$ представляет собой возможную последовательность выравнивания предложений при условии, что $i(t)\leq i(t+1)$ и $j(t)\leq j(t+1)$.

В ходе поиска наилучшей последовательности выравнивания $(i(t),j(t))^*$ методом динамического программирования используются два правила:

- пара предложений $(e_{i(t)},d_{j(t)})$ может быть включена в последовательность выравнивания при одновременном выполнении двух условий:
$$\max\{Tr(e_{i(t)},d_{j(t)}),Tr(d_{j(t)}|e_{i(t)})\}>Tr_{\max}$$
и
$$\min\{Tr(d_{j(t)}|e_{i(t)}), Tr(e_{i(t)},d_{j(t)})\}>Tr_{\min},$$
где $Tr_{\max}$ и $Tr_{\min}$ – эмпирически подбираемые параметры, в нашем случае – 0.5 и 0.25 соответственно. Увеличение значений $Tr_{\max}$ и $Tr_{\min}$ приводит к повышению точности выравнивания, а их уменьшение – к повышению полноты за счет снижения точности. Чем больше полнота используемого словаря переводных соответствий, тем более высокими могут быть выбраны значения $Tr_{\max}$ и $Tr_{\min}$;
- последовательность выравнивания A признается лучше другой последовательности B, если величина $\sum_t(Eq(e_{i(t)}|,d_{j(t)})+Eq(d_{j(t)}|\ e_{i(t)}))$ – совокупное количество сопоставлений переводов – для последовательности A превышает таковую величину для последовательности B.

После нахождения наилучшего отображения параллельных предложений «одно к одному» делается попытка отобразить предложения, пропущенные в последовательности выравнивания, на те предложения, с которыми уже выровнены предложения, соседние с пропущенными, при реализации выравнивания «одно к нескольким» для случаев несинхронной разбивки исходного и целевого текста на предложения. В контексте задачи обучения СРС процедура выравнивания имеет целью получить смысло-текст минимального размера, поэтому разрешается объединять в один смысло-текст не более двух предложений. В финале происходит отбрасывание тех пар смысло-текстов, для которых мера прямого или обратного сходства переводов оказывается ниже определенного порога –

ожидается, что соответствующий перевод является излишне «вольным».

4 Реализация и эксперименты

Эксперименты по обучению СРС были проведены на корпусе новостных текстов, полученных с армянского сайта http://news.am. Из двух разделов данного сайта (http://news.am/rus/news/ и http://news. am/arm/news/) были скачаны по 300 тысяч русских и армянских текстов, из числа которых по формальному признаку – совпадению идентификаторов – было получено 230 тысяч пар предположительно параллельных русско-армянских текстов.

Для анализа русских текстов был использован лингвистический анализатор RCO Fact Extractor [8], который проводил полный синтаксический анализ текста, выделяя сущности разных типов с отношениями между ними, а также события и факты с их участниками в соответствии с заданными синтактико-семантическими шаблонами [9]. Для анализа армянских текстов был разработан неполный лингвистический анализатор, который разбивал текст на слова и предложения, проводил морфологический анализ и определял для каждого слова возможные варианты его нормальной формы, а также распознавал на основе формальных правил и сворачивал в одну сущность особые цепочки слов – обозначения именованных персон, организаций, географических объектов, дат и обстоятельств времени. Основой для построения армянского морфословаря послужил Восточно-армянский национальный корпус [1], правила описания особых сущностей были разработаны лингвистом на языке Cape для компонента RCO Pattern Extractor [10].

Армяно-русский словарь переводов содержал более 100 тысяч единиц и был сформирован путем консолидации переводов из нескольких интернет-источников. Статистические переводчики Яндекс и Гугл могут переводить по-разному различные словоформы одного и того же слова, например, разным формам армянского слова «ծառայություն» соответствуют формы русских слов *сервис, служба, услуга, обслуживание*, а также ряд ошибочных переводов. Эмпирически было подобрано правило определения достоверности переводов, согласно которому признаются недостоверными те варианты, которые встречаются со взвешенной частотой, отношение которой к взвешенной частоте самого частого варианта составляет менее 0.7. Взвешенная частота есть сумма частот встречаемости в каждом из источников, умноженных на вес источника, который определяет уровень доверия к нему. На практике были использованы три источника переводов: а) переводы встретившихся в текстах словоформ, полученные из Яндекса, с весом 1; б) переводы тех же словоформ, полученные из Гугла, с весом 2 (переводы Гугла мы считали достовернее переводов Яндекса); в) строгий словарь объемом 22 тысячи слов (нормальных форм), полученный из интернет-источника http://www.classes.ru/all-armenian/dictionary-armenian-russian.htm, с весом 100, что означало отброс всех вариантов Яндекс- и Гугл-переводов слов, встретившихся в строгом словаре. На переводы в Яндекс и Google были отправлены все армянские словоформы, встретившихся не менее чем в двух документах 230-тысячного корпуса текстов, а также именованные сущности, что составило 350 тысяч единиц.

С использованием полученного словаря переводов алгоритм, описанный в Разделе **3**, разбил 230 тысяч пар текстов на 1370 тысяч пар параллельных фрагментов – смысло-текстов, а для 690 тысяч русских и 585 тысяч армянских предложений не было найдено достаточно близких параллельных переводов. Данная процедура заняла около восьми часов работы одного процессорного ядра.

Программные компоненты обучения СРС работают в три фазы.

На Фазе I обрабатывается корпус xml-файлов, которые формируются двумя лингвистическими анализаторами и содержат описание сущностей, выделенных в армянских смысло-текстах, а также идентификаторы смыслов, которым релевантны параллельные им русские смысло-тексты. Собираются все *n*-ки из нормальных форм сущностей, упоминавшиеся в армянских смысло-текстах, длиной от 2 до 4, допуская встречаемость между словами *n*-ок посторонних слов количеством от 0 до 5. Также собираются параметризованные варианты *n*-ок, в которых конкретные именованные сущности заменяются на свои типы – персона, организация, география. Все омонимичные варианты нормальных форм сущностей порождают соответствующие варианты *n*-ок. Количество разных *n*-ок, получаемых таким образом, имеет порядок сотен миллионов, поэтому для хранения статистики (общие частоты встречаемости *n*-ок в корпусе и частоты *n*-ок по каждому смыслу) в оперативной памяти применяется процедура периодического забывания – как только количество сохраненных *n*-ок превышает 10 миллионов (что не превышает 2 Гбайт ОЗУ), из памяти удаляются данные по наиболее редко встретившимся *n*-кам, имеющим низкие оценки правдоподобия вхождения в профиль какого-либо смысла в соответствии с (3). В финале для каждого смысла отбирается до 1,5 тысяч лучших *n*-ок – кандидатов на последующее включение в профиль, получивших наибольшие оценки правдоподобия вхождения в профиль $q(s_j)$ в соответствии с (3), но не менее 0.01, и сохраняются в файле – препрофиле смысла. Время обработки 230 тысяч новостных текстов для 40 смыслов (см. Таблицу 1) на этой фазе занимает около 4 часов работы одного процессорного ядра.

На Фазе II загружаются файлы препрофилей смыслов, и вновь обрабатывается корпус xml-файлов с описаниями сущностей, выделенных в параллельных смысло-текстах. В результате для каждой *n*-ки в препрофилях подсчитываются частоты ее встречаемости в окнах различной длины с количеством допустимых вставок посторонних слов

от 0 до 5. Одновременно для *n*-ки собираются идентифкаторы смысло-текстов, ее содержащих, по каждому из окон. Собранная информация сохраняется в полных файлах препрофилей смыслов. Время выполнения этой фазы составляет около 1 часа.

На Фазе III загружается файл с полной информацией об *n*-ках препрофилей смыслов и выполняется алгоритм построения профиля СРС, который выбирает *n*-ки из препрофиля в профиль, вычисляя для каждой возможной комбинации n-ок оценку правдоподобия и запоминая комбинацию с максимальной оценкой как лучший вариант профиля S^* в соответствии с (2). Максимальное количество просматриваемых комбинаций ограничивалось 1 миллионом, что оказалось с избытком достаточно для получения наилучшего варианта профиля – средний номер шага процедуры перебора комбинаций, на котором был получен наилучший вариант S^*, по 40 профилям составил около 2 тысяч, а наибольшее из значений (для профиля «путешествия») не превышает 20 тысяч. Для большинства смыслов количество всех комбинаций, подлежащих проверке на выполнение условий (4), оказалось значительно меньше миллиона вследствие относительно небольшого количества обучающих примеров и соответствующих *n*-ок-кандидатов на включение в профиль. В итоге время выполнения данной фазы составило в среднем одну секунду на профиль.

Настройка СРС проводилась на полученном корпусе из 1.370 тысяч пар параллельных смысло-текстов для 40 смыслов – ситуаций, отобранных из более чем 200 типовых ситуаций, распознаваемых русскоязычными лингвистическими шаблонами RCO Fact Extractor. Названия этих смыслов-ситуаций приведены в первом столбце Таблицы 1. Именно к ним обнаружено в корпусе наибольшее количество релевантных смысло-текстов, которое указано в третьем столбце *Exm*.

В экспериментах было построено два отдельных СРС – профили русского СРС строились на русских смысло-текстах и состояли из *n*-ок русских сущностей, выделенных RCO Fact Exctractor, а профили армянского СРС строились на параллельных армянских смысло-текстах и состояли из n-ок армянских слов, выделенных разработанным армянским лингвоанализатором. Обучение СРС «с русского на русский» позволяло исследовать работу СРС в чистом виде, без влияния факторов посторонних составляющих– несовершенств армянского лингвоанализатора, процедуры выравнивания параллельных фрагментов и недостатков собственно параллельных переводов. Значения, полученные для армянского и русского СРС, в Таблице 1 приведены вместе и разделены символом '/'.

Каждая из 40 ситуаций предполагает вовлечение в нее одного или двух участников, представленных в тексте произвольными именованными сущностями. Поэтому, вместо конкретных слов – имен собственных, *n*-ки профилей включали в себя обозначения типов именованных сущностей (O – организация, P – персона, G – географическое место), которые указаны во втором столбце Таблицы 1. Знак '|' разделяет возможные альтернативы. Например, для смысла *владение акциями* во втором столбце указано *O|P O*, что означает, что в *n*-ку слов, входящую в профиль данного смысла, должны обязательно войти какая-либо именованная персона или организация (*владелец акций*) плюс именованная организация (*эмитент акций*).

Различия между цифрами (количество релевантных смысло-текстов) в столбцах *Exm* и *TrainExm* обусловлено следующим. Оценка правдоподобия вхождения *n*-ки в профиль смысла $q(s_j)$ в соответствии с (3) равна 0 в случае единичной частоты встречаемости *n*-ки, вследствие чего такие *n*-ки не могли быть включены в профиль в силу объективной недостаточности данных для обучения СРС. В результате этого многие смысло-тексты, не содержащие ни одной *n*-ки с частотой более 1 и относительно высоким значением $q(s_j)>0{,}01$, фактически не могли участвовать в обучении. Поэтому при расчете значений *R* в соответствии с (1) в качестве $D(Se)$ бралось множество смысло-текстов, содержащих хотя бы одну из *n*-ок-кандидатов на включение в профиль. Это позволяло оценить качество алгоритма обучения относительно независимо от качества обучающей выборки, а также от качества лингвистического анализа армянского текста, которое априори было хуже качества анализа русского – прежде всего, экспериментальный морфоанализатор для армянского языка не мог приводить разные формы слова к одной форме с такой же полнотой и точностью, как морфоанализатор для русского языка. Именно эти фактором в первую очередь обусловлено то, что среднее по столбцу *TrainExm* для армянского языка – 509 – оказалось вдвое меньше, чем для русского – 1056. Соответственно, среднее количество *n*-грамм, включенных в армянские профили, в столбце *n-s* – 145 – оказалось меньше, чем для русского языка – 173. Кроме того, армянские переводы русских новостных текстов нередко опускают описания деталей событий, в которых содержится целевой смысл в исходных текстах, распознаваемый русским лингвоанализатором.

С учетом сказанного средние значения полноты (0.61 для армянского языка против 0.71 для русского в столбце *R*) и точности (0.94 для армянского языка против 0.91 для русского) представляются нам близкими. Соответствующие значения F_1-меры, балансирующей полноту и точность в соответствии с (1), различаются еще меньше – 0.73 против 0.78. Реально ожидаемая полнота, рассчитанная с учетом всех 3587 примеров в корпусе, для русских текстов составляет около 0,21 (0.71 умножить на 1056/3587), а для армянских текстов – 0,09 (0.61 умножить на 509/3587).

Таблица 1 Данные по профилям смыслов. Имена столбцов: Sense – имя смысла; Param – типы сущностей-параметров в *n*-ках; Exm – количество релевантных смысло-текстов в обучающем корпусе; TrainExm – количество релевантных смысло-текстов, участвовавших в обучении профиля; *n–s* – количество *n*-ок, вошедших в профиль; P, R – точность и полнота на обучающем корпусе в соответствии с (1). Символом '/' разделены значения, полученные на армянских и русских смысло-текстах

Sense	**Param**	**Exm**	**TrainExm**	**n-s**	**P**	**R**
митинги/забастовки	G	3351	1205/2119	500/500	0.95/0.87	0.62/0.67
уход с рынка	O	25	13/19	5/6	1.0/1.0	0.92/0.74
поставки	O	130	18/71	8/34	0.92/0.90	0.61/0.76
предоставление услуг	O	70	21/43	4/18	1.0/0.88	0.33/0.65
открытие торг. точек	O	224	26/20	11/9	1.0/0.86	0.58/0.95
новые проекты	O	88	34/49	15/21	0.95/0.94	0.56/0.65
проведение тендера	O	108	37/79	10/39	0.96/0.92	0.59/0.82
отзыв продукции	O	131	50/63	23/30	0.94/0.94	0.68/0.78
открытие филиала	O	167	79/122	33/50	0.96/0.88	0.63/0.82
купля/продажа акций	O	437	153/252	58/83	0.98/0.86	0.64/0.88
выпуск товаров	O	549	192/333	61/102	1.0/0.92	0.39/0.52
создание компании	O	565	213/185	51/59	0.98/0.98	0.49/0.68
экономические показатели	O	3212	346/857	182/178	0.97/0.92	0.85/0.69
объединение	O O	222	14/40	1/12	0.83/0.82	0.36/0.80
партнерство	O O	479	79/143	23/53	0.87/0.95	0.58/0.57
рейтинги	O|P	165	29/89	10/29	0.89/0.84	0.55/0.73
юбилей	O|P	90	30/69	10/26	0.89/0.84	0.53/0.86
банкротство	O|P	114	46/70	14/26	0.89/0.92	0.54/0.79
купля/продажа финансов	O|P	750	100/123	36/66	0.93/0.93	0.57/0.78
выигрыш призов	O|P	583	255/374	107/132	0.99/0.91	0.54/0.71
благотворительность	O|P	604	257/285	112/98	0.98/0.91	0.63/0.76
скандалы	O|P	6895	511/1699	166/284	0.94/0.86	0.66/0.65
суды, расследования	O|P	4657	643/2123	93/192	0.95/0.97	0.73/0.60
конфликты	O|P O|P	9932	647/3510	109/380	0.96/0.90	0.72/0.63
финансовая деятельность	O|P	5939	691/1083	230/159	0.94/0.90	0.67/0.72
успехи–неудачи	O|P	5899	1093/2141	331/355	0.89/0.90	0.63/0.72
планы/намерения	O|P	7948	1374/2055	390/273	0.93/0.91	0.54/0.59
мероприятия	O|P	23698	2867/6575	500/500	0.92/0.92	0.61/0.65
владение акциями	O|P O	365	40/95	14/42	0.91/0.99	0.53/0.71
владение организациями	O|P O	2040	431/679	135/183	0.87/0.84	0.69/0.78
договора	O|P O|P	6252	595/1020	211/313	0.93/0.93	0.58/0.67
отставка с должности	P	941	355/521	166/143	0.95/0.92	0.79/0.84
авторство	P	1740	383/555	150/126	0.96/0.92	0.55/0.73
кандидат на выборах	P	2068	529/1020	217/279	0.95/0.91	0.65/0.68
письма	P	2951	819/1102	229/216	0.88/0.87	0.68/0.80
назначение на должность	P	5458	1192/2449	352/417	0.88/0.90	0.72/0.61
путешествия	P G	13295	2564/4304	500/500	0.87/0.86	0.45/0.66
физическое насилие	P P	292	56/142	11/47	1.0/0.95	0.68/0.73
разговор	P P	10216	889/2019	299/436	0.93/0.92	0.55/0.64
встреча	P P	20838	1494/3730	429/500	0.89/0.84	0.73/0.56
среднее		**3587**	**509/1056**	**145/173**	**0.94/0.91**	**0.61/0.71**

5 Заключение

Предложен и экспериментально исследован подход к распознаванию смыслов (упоминаний целевых ситуаций, событий и фактов) в тексте, который допускает относительно простую реализацию предположительно для любого языка, при наличии возможности автоматического выделения требуемых смыслов на русском языке. Подход требует наличия корпуса квазипараллельных текстов – переводов с русского языка на иностранный или обратно. Также желательно наличие простейшего лингвистического анализатора, способного строить варианты нормальных форм для словоформ иностранного языка, что позволяет существенно повысить полноту распознавания смыслов, не требуя примеров параллельных текстов, в которых описывающие смысл слова стоят во всех возможных формах. В зависимости от видов распознаваемых смыслов от лингвистического анализатора может потребоваться умение выделять именованные сущности.

Описанные эксперименты показали высокую точность распознавания смыслов для большого количества разнообразных смыслов (40) на обучающей выборке большого объема (230 тысяч пар квазипараллельных текстов, более 1370 тысяч пар армянских и русских предложений), что, в силу особенностей выбранного способа описания смысла (*n*-ок слов, совместно встречающихся в окне), позволяет ожидать высокой точности распознавания и на других текстах. Невысокая полнота распознавания говорит о необходимости увеличить размер корпуса параллельных новостных текстов в несколько раз (с 230 тысяч пар до миллиона).

В экспериментах не использовалась контрольная выборка текстов, отличная от обучающей, для проверки полученных оценок ожидаемой точности и полноты в силу отсутствия возможности получения качественной экспертной разметки корпуса не только армянских, но и каких-либо других текстов на предмет релевантности различным смыслам. Тем не менее, просмотр содержимого построенных профилей – русских и армянских *n*-ок слов – показал релевантность подавляющего большинства из них целевым смыслам, что повышает уверенность в эффективности подхода.

Литература

[1] Eastern Armenian National Corpus, http://eanc.net

[2] Grefenstette, G. (ed.): Cross-Language Information Retrieval. Springer, 177 p. (1998)

[3] He, D., Wang, J.: Cross-Language Information Retrieval. Information Retrieval: Searching in the 21st Century, Part 11. Wiley and Sons Ltd, pp. 233-254 (2009)

[4] Nie, J-Y.: Cross-Language Information Retrieval. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers, 3 (1), pp. 1-125 (2010)

[5] Nie, J-Y., Gao, J., Cao, G.: Translingual Mining from Text Data. Mining Text Data, Part X. Springer US, pp. 323-359 (2012)

[6] SMT Research Survey Wiki: A Comprehensive Survey of Statistical Machine Translation Research Publications. Sentence Alignment, http://www.statmt.org/survey/Topic/SentenceAlignment

[7] Statistical Machine Translation, maintained by Philipp Koehn, http://www.statmt.org

[8] RCO Fact Extractor – инструмент компьютерного анализа текстовой информации компании «ЭР СИ О», http://www.rco.ru/?page_id=3554

[9] Ермаков, А.Е., Плешко, В.В.: Семантическая интерпретация в системах компьютерного анализа текста. Информационные технологии, (6), сс. 2-7 (2009)

[10] Ермаков, А.Е., Плешко, В.В., Митюнин, В.А.: RCO Pattern Extractor: компонент выделения особых объектов в тексте. Информатизация и информационная безопасность правоохранительных органов: Сборник трудов XII Межд. науч. конф., Москва, сс. 312-317 (2003)

[11] Потемкин, С.Б., Кедрова, Г.Е.: Выравнивание неразмеченного корпуса параллельных текстов. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Межд. конф. «Диалог» (Бекасово, 4–8 июня 2008 г.). Вып. 7 (14). М.: РГГУ, сс. 431-437 (2008)